Chapter 2: Graphical Descriptions of Data

In chapter 1, you were introduced to the concepts of population, which again is a collection of all the measurements from the individuals of interest. Remember, in most cases you can't collect the entire population, so you have to take a sample. Thus, you collect data either through a sample or a census. Now you have a large number of data values. What can you do with them? No one likes to look at just a set of numbers. One thing is to organize the data into a table or graph. Ultimately though, you want to be able to use that graph to interpret the data, to describe the distribution of the data set, and to explore different characteristics of the data. The characteristics that will be discussed in this chapter and the next chapter are:

1. Center: middle of the data set, also known as the average.
2. Variation: how much the data varies.
3. Distribution: shape of the data (symmetric, uniform, or skewed).
4. Qualitative data: analysis of the data
5. Outliers: data values that are far from the majority of the data.
6. Time: changing characteristics of the data over time.

This chapter will focus mostly on using the graphs to understand aspects of the data, and not as much on how to create the graphs. There is technology that will create most of the graphs, though it is important for you to understand the basics of how to create them.

Section 2.1: Qualitative Data

Remember, qualitative data are words describing a characteristic of the individual. There are several different graphs that are used for qualitative data. These graphs include bar graphs, Pareto charts, and pie charts.

Pie charts and bar graphs are the most common ways of displaying qualitative data. A spreadsheet program like Excel can make both of them. The first step for either graph is to make a **frequency or relative frequency table**. A frequency table is a summary of the data with counts of how often a data value (or category) occurs.

Example #2.1.1: Creating a Frequency Table

Suppose you have the following data for which type of car students at a college drive?

Ford, Chevy, Honda, Toyota, Toyota, Nissan, Kia, Nissan, Chevy, Toyota, Honda, Chevy, Toyota, Nissan, Ford, Toyota, Nissan, Mercedes, Chevy, Ford, Nissan, Toyota, Nissan, Ford, Chevy, Toyota, Nissan, Honda, Porsche, Hyundai, Chevy, Chevy, Honda, Toyota, Chevy, Ford, Nissan, Toyota, Chevy, Honda, Chevy, Saturn, Toyota, Chevy, Chevy, Nissan, Honda, Toyota, Toyota, Nissan

A listing of data is too hard to look at and analyze, so you need to summarize it. First you need to decide the categories. In this case it is relatively easy; just use the car type. However, there are several cars that only have one car in the list. In that case it is easier to make a category called other for the ones with low values. Now just count how many of each type of cars there are. For example, there are 5 Fords, 12 Chevys, and 6 Hondas. This can be put in a frequency distribution:

Table #2.1.1: Frequency Table for Type of Car Data

Category	Frequency
Ford	5
Chevy	12
Honda	6
Toyota	12
Nissan	10
Other	5
Total	50

The total of the frequency column should be the number of observations in the data.

Since raw numbers are not as useful to tell other people it is better to create a third column that gives the relative frequency of each category. This is just the frequency divided by the total. As an example for Ford category:

$$\text{relative frequency} = \frac{5}{50} = 0.10$$

This can be written as a decimal, fraction, or percent. You now have a relative frequency distribution:

Table #2.1.2: Relative Frequency Table for Type of Car Data

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

The relative frequency column should add up to 1.00. It might be off a little due to rounding errors.

Now that you have the frequency and relative frequency table, it would be good to display this data using a graph. There are several different types of graphs that can be used: bar chart, pie chart, and Pareto charts.

Bar graphs or charts consist of the frequencies on one axis and the categories on the other axis. Then you draw rectangles for each category with a height (if frequency is on the vertical axis) or length (if frequency is on the horizontal axis) that is equal to the frequency. All of the rectangles should be the same width, and there should be equally width gaps between each bar.

Example #2.1.2: Drawing a Bar Graph

Draw a bar graph of the data in example #2.1.1.

Table #2.1.2: Frequency Table for Type of Car Data

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

Put the frequency on the vertical axis and the category on the horizontal axis. Then just draw a box above each category whose height is the frequency.

All graphs are drawn using R. The command in R to create a bar graph is:

variable<-c(type in percentages or frequencies for each class with commas in between values)

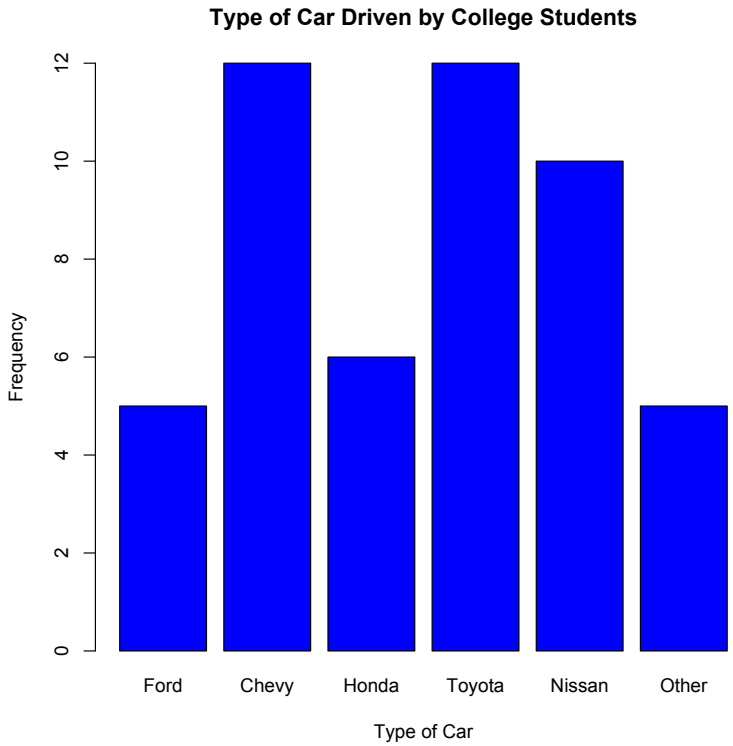
barplot(variable,names.arg=c("type in name of 1st category", "type in name of 2nd category",…,"type in name of last category"), ylim=c(0,number over max), xlab="type in label for x-axis", ylab="type in label for y-axis",ylim=c(0,number above maximum y value), main="type in title", col="type in a color") – creates a bar graph of the data in a color if you want.

For this example the command would be:

car<-c(5, 12, 6, 12, 10, 5)

barplot(car, names.arg=c("Ford", "Chevy", "Honda", "Toyota", "Nissan", "Other"), xlab="Type of Car", ylab="Frequency", ylim=c(0,12), main="Type of Car Driven by College Students", col="blue")

Graph #2.1.1: Bar Graph for Type of Car Data



Notice from the graph, you can see that Toyota and Chevy are the more popular car, with Nissan not far behind. Ford seems to be the type of car that you can tell was the least liked, though the cars in the other category would be liked less than a Ford.

Some key features of a bar graph:

- Equal spacing on each axis.
- Bars are the same width.
- There should be labels on each axis and a title for the graph.
- There should be a scaling on the frequency axis and the categories should be listed on the category axis.
- The bars don't touch.

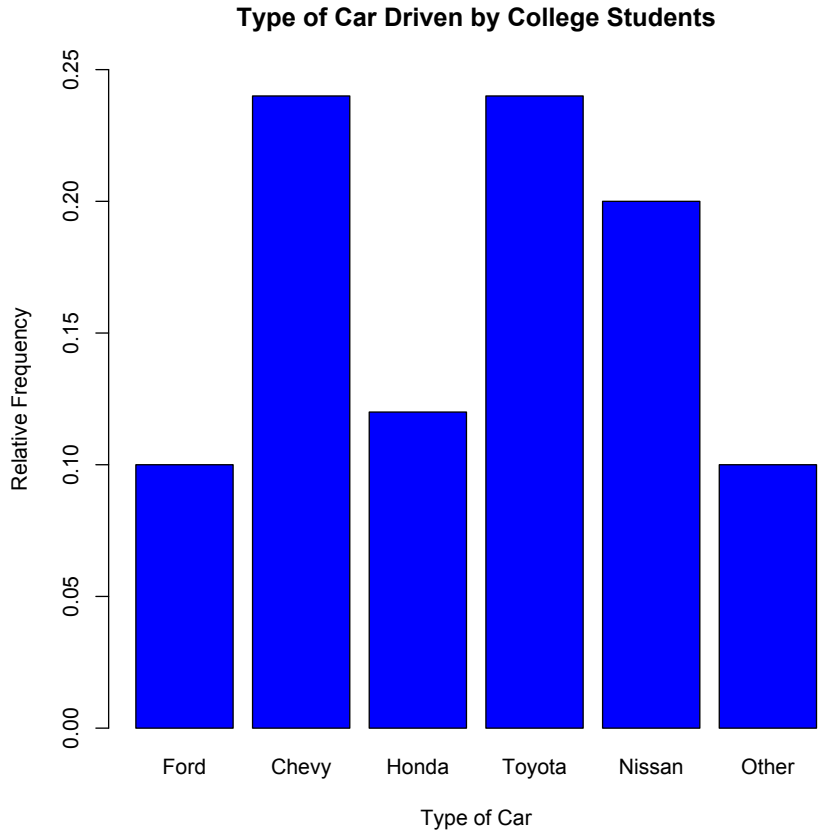
You can also draw a bar graph using relative frequency on the vertical axis. This is useful when you want to compare two samples with different sample sizes. The relative frequency graph and the frequency graph should look the same, except for the scaling on the frequency axis.

Using R, the command would be:

car<-c(0.1, 0.24, 0.12, 0.24, 0.2, 0.1)

barplot(car, names.arg=c("Ford", "Chevy", "Honda", "Toyota", "Nissan", "Other"), xlab="Type of Car", ylab="Relative Frequency", main="Type of Car Driven by College Students", col="blue", ylim=c(0,.25))

Graph #2.1.2: Relative Frequency Bar Graph for Type of Car Data



Another type of graph for qualitative data is a pie chart. A pie chart is where you have a circle and you divide pieces of the circle into pie shapes that are proportional to the size of the relative frequency. There are 360 degrees in a full circle. Relative frequency is just the percentage as a decimal. All you have to do to find the angle by multiplying the relative frequency by 360 degrees. Remember that 180 degrees is half a circle and 90 degrees is a quarter of a circle.

Example #2.1.3: Drawing a Pie Chart

Draw a pie chart of the data in example #2.1.1.

First you need the relative frequencies.

Table #2.1.2: Frequency Table for Type of Car Data

Category	Frequency	Relative Frequency
Ford	5	0.10
Chevy	12	0.24
Honda	6	0.12
Toyota	12	0.24
Nissan	10	0.20
Other	5	0.10
Total	50	1.00

Then you multiply each relative frequency by 360° to obtain the angle measure for each category.

Table #2.1.3: Pie Chart Angles for Type of Car Data

Category	Relative Frequency	Angle (in degrees (°))
Ford	0.10	36.0
Chevy	0.24	86.4
Honda	0.12	43.2
Toyota	0.24	86.4
Nissan	0.20	72.0
Other	0.10	36.0
Total	1.00	360.0

Now draw the pie chart using a compass, protractor, and straight edge. Technology is preferred. If you use technology, there is no need for the relative frequencies or the angles.

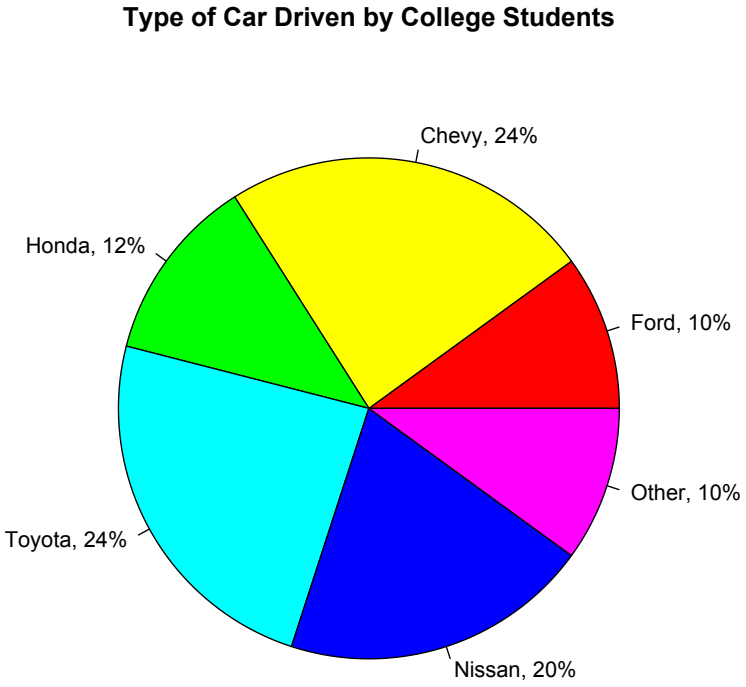
You can use R to graph the pie chart. In R, the commands would be:

pie(variable,labels=c("type in name of 1st category", "type in name of 2nd category",…,"type in name of last category"),main="type in title", col=rainbow(number of categories)) – creates a pie chart with a title and rainbow of colors for each category.

For this example, the commands would be:

```
car<-c(5, 12, 6, 12, 10, 5)
pie(car, labels=c("Ford, 10%", "Chevy, 24%", "Honda, 12%", "Toyota, 24%", "Nissan, 20%", "Other, 10%"), main="Type of Car Driven by College Students", col=rainbow(6))
```

Graph #2.1.3: Pie Chart for Type of Car Data

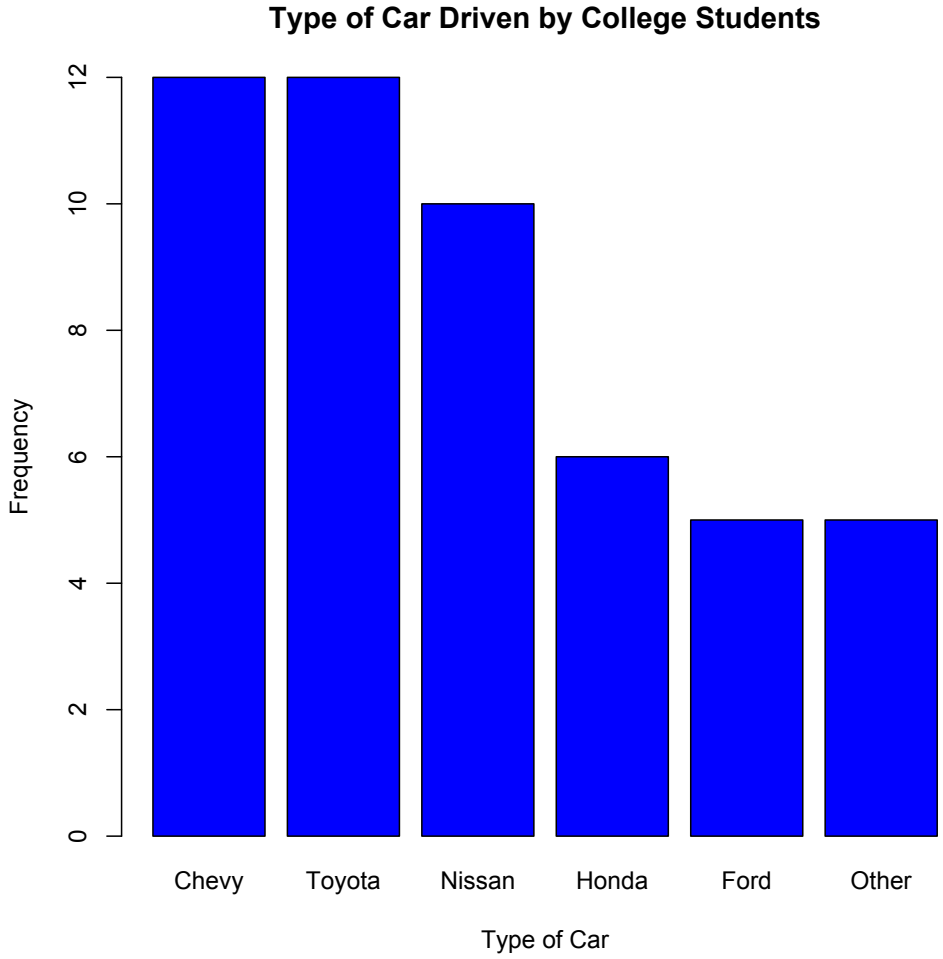


As you can see from the graph, Toyota and Chevy are more popular, while the cars in the other category are liked the least. Of the cars that you can determine from the graph, Ford is liked less than the others.

Pie charts are useful for comparing sizes of categories. Bar charts show similar information. It really doesn't matter which one you use. It really is a personal preference and also what information you are trying to address. However, pie charts are best when you only have a few categories and the data can be expressed as a percentage. The data doesn't have to be percentages to draw the pie chart, but if a data value can fit into multiple categories, you cannot use a pie chart. As an example, if you asking people about what their favorite national park is, and you say to pick the top three choices, then the total number of answers can add up to more than 100% of the people involved. So you cannot use a pie chart to display the favorite national park.

A third type of qualitative data graph is a **Pareto chart,** which is just a bar chart with the bars sorted with the highest frequencies on the left. Here is the Pareto chart for the data in Example #2.1.1.

Graph #2.1.4: Pareto Chart for Type of Car Data



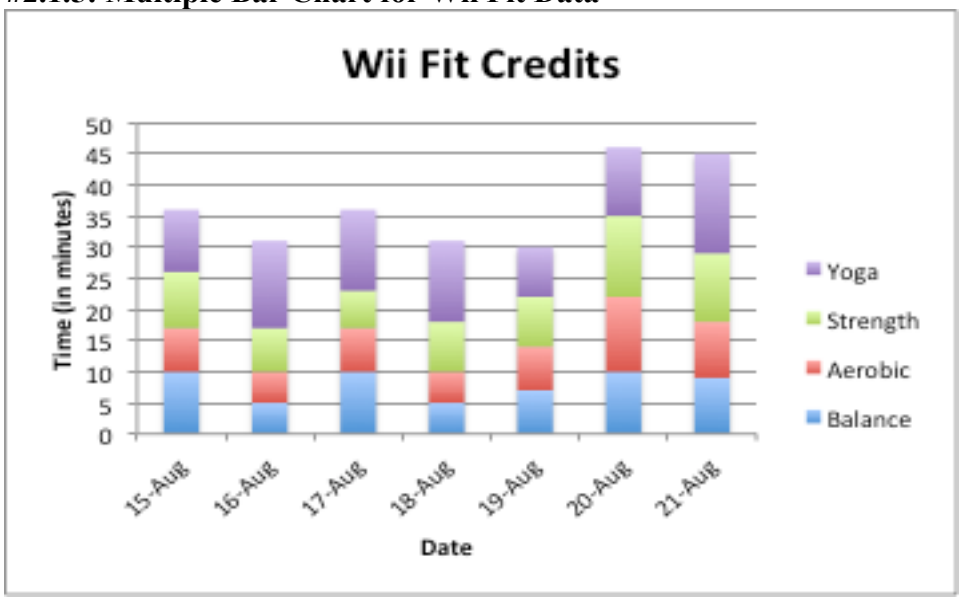
The advantage of Pareto charts is that you can visually see the more popular answer to the least popular. This is especially useful in business applications, where you want to know what services your customers like the most, what processes result in more injuries, which issues employees find more important, and other type of questions like these.

There are many other types of graphs that can be used on qualitative data. There are spreadsheet software packages that will create most of them, and it is better to look at them to see what can be done. It depends on your data as to which may be useful. The next example illustrates one of these types known as a multiple bar graph.

Example #2.1.4: Multiple Bar Graph

In the Wii Fit game, you can do four different types if exercises: yoga, strength, aerobic, and balance. The Wii system keeps track of how many minutes you spend on each of the exercises everyday. The following graph is the data for Dylan over one week time period. Discuss any indication you can infer from the graph.

Graph #2.1.5: Multiple Bar Chart for Wii Fit Data



Solution:
It appears that Dylan spends more time on balance exercises than on any other exercises on any given day. He seems to spend less time on strength exercises on a given day. There are several days when the amount of exercise in the different categories is almost equal.

The usefulness of a multiple bar graph is the ability to compare several different categories over another variable, in example #2.1.4 the variable would be time. This allows a person to interpret the data with a little more ease.

Section 2.1: Homework

1.) Eyeglassomatic manufactures eyeglasses for different retailers. The number of lenses for different activities is in table #2.1.4.

Table #2.1.4: Data for Eyeglassomatic

Activity	Grind	Multicoat	Assemble	Make frames	Receive finished	Unknown
Number of lenses	18872	12105	4333	25880	26991	1508

Grind means that they ground the lenses and put them in frames, multicoat means that they put tinting or scratch resistance coatings on lenses and then put them in frames, assemble means that they receive frames and lenses from other sources and put them together, make frames means that they make the frames and put lenses in from other sources, receive finished means that they received glasses from other source, and unknown means they do not know where the lenses came from. Make a bar chart and a pie chart of this data. State any findings you can see from the graphs.

2.) To analyze how Arizona workers ages 16 or older travel to work the percentage of workers using carpool, private vehicle (alone), and public transportation was collected. Create a bar chart and pie chart of the data in table #2.1.5. State any findings you can see from the graphs.

Table #2.1.5: Data of Travel Mode for Arizona Workers

Transportation type	Percentage
Carpool	11.6%
Private Vehicle (Alone)	75.8%
Public Transportation	2.0%
Other	10.6%

3.) The number of deaths in the US due to carbon monoxide (CO) poisoning from generators from the years 1999 to 2011 are in table #2.1.6 (Hinatov, 2012). Create a bar chart and pie chart of this data. State any findings you see from the graphs.

Table #2.1.6: Data of Number of Deaths Due to CO Poisoning

Region	Number of deaths from CO while using a generator
Urban Core	401
Sub-Urban	97
Large Rural	86
Small Rural/Isolated	111

4.) In Connecticut households use gas, fuel oil, or electricity as a heating source. Table #2.1.7 shows the percentage of households that use one of these as their principle heating sources ("Electricity usage," 2013), ("Fuel oil usage," 2013), ("Gas usage," 2013). Create a bar chart and pie chart of this data. State any findings you see from the graphs.

Table #2.1.7: Data of Household Heating Sources

Heating Source	Percentage
Electricity	15.3%
Fuel Oil	46.3%
Gas	35.6%
Other	2.8%

5.) Eyeglassomatic manufactures eyeglasses for different retailers. They test to see how many defective lenses they made during the time period of January 1 to March 31. Table #2.1.8 gives the defect and the number of defects. Create a Pareto chart of the data and then describe what this tells you about what causes the most defects.

Table #2.1.8: Data of Defect Type

Defect type	Number of defects
Scratch	5865
Right shaped – small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape – big	1105
Lost in lab	976
Spots/bubble – intern	976

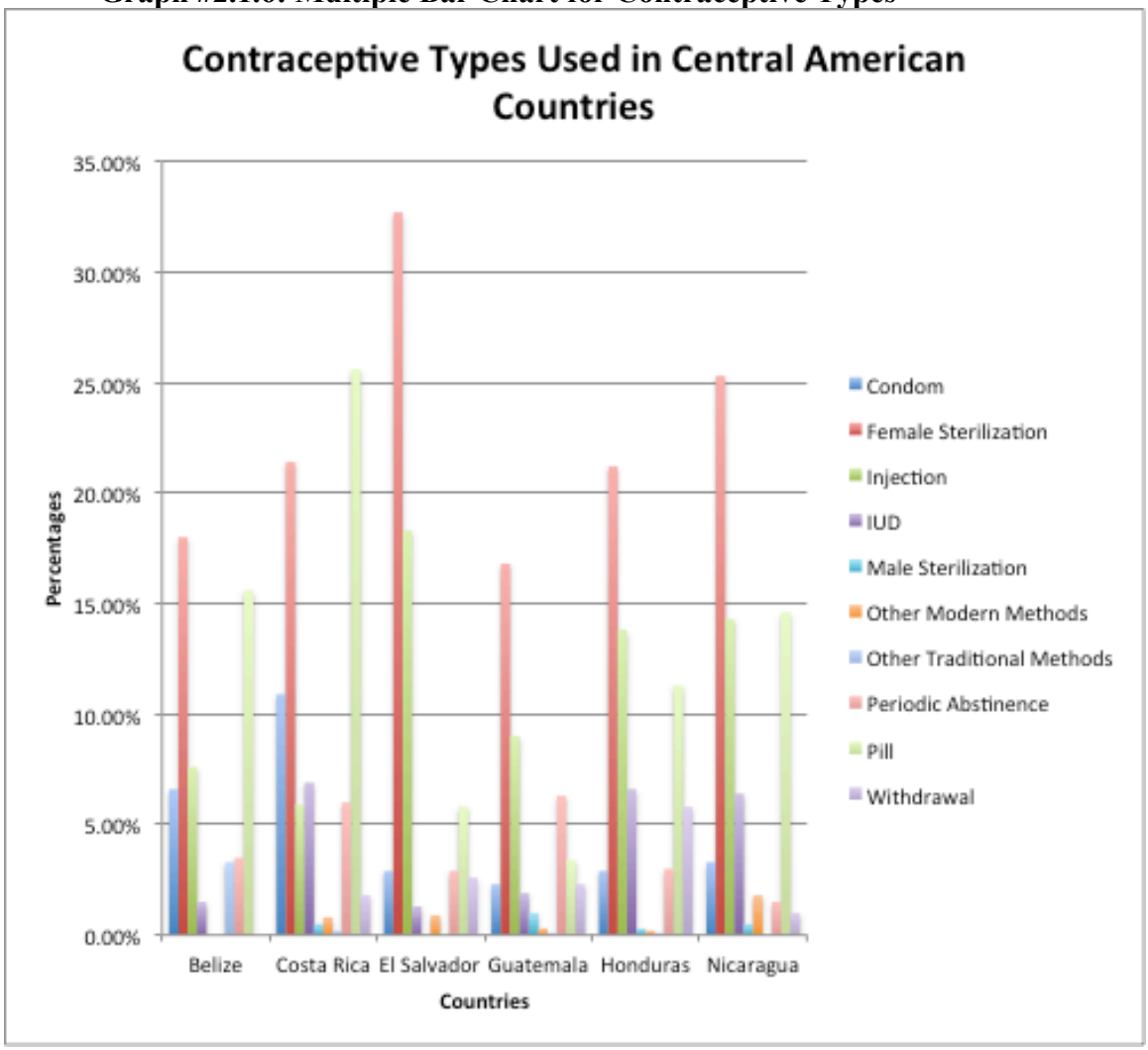
6.) People in Bangladesh were asked to state what type of birth control method they use. The percentages are given in table #2.1.9 ("Contraceptive use," 2013). Create a Pareto chart of the data and then state any findings you can from the graph.

Table #2.1.9: Data of Birth Control Type

Method	Percentage
Condom	4.50%
Pill	28.50%
Periodic Abstinence	4.90%
Injection	7.00%
Female Sterilization	5.00%
IUD	0.90%
Male Sterilization	0.70%
Withdrawal	2.90%
Other Modern Methods	0.70%
Other Traditional Methods	0.60%

7.) The percentages of people who use certain contraceptives in Central American countries are displayed in graph #2.1.6 ("Contraceptive use," 2013). State any findings you can from the graph.

Graph #2.1.6: Multiple Bar Chart for Contraceptive Types



Section 2.2: Quantitative Data

The graph for quantitative data looks similar to a bar graph, except there are some major differences. First, in a bar graph the categories can be put in any order on the horizontal axis. There is no set order for these data values. You can't say how the data is distributed based on the shape, since the shape can change just by putting the categories in different orders. With quantitative data, the data are in specific orders, since you are dealing with numbers. With quantitative data, you can talk about a distribution, since the shape only changes a little bit depending on how many categories you set up. This is called a **frequency distribution**.

This leads to the second difference from bar graphs. In a bar graph, the categories that you made in the frequency table were determined by you. In quantitative data, the categories are numerical categories, and the numbers are determined by how many categories (or what are called classes) you choose. If two people have the same number of categories, then they will have the same frequency distribution. Whereas in qualitative data, there can be many different categories depending on the point of view of the author.

The third difference is that the categories touch with quantitative data, and there will be no gaps in the graph. The reason that bar graphs have gaps is to show that the categories do not continue on, like they do in quantitative data. Since the graph for quantitative data is different from qualitative data, it is given a new name. The name of the graph is a **histogram**. To create a histogram, you must first create the frequency distribution. The idea of a frequency distribution is to take the interval that the data spans and divide it up into equal subintervals called classes.

Summary of the steps involved in making a frequency distribution:

1. Find the range = largest value – smallest value
2. Pick the number of classes to use. Usually the number of classes is between five and twenty. Five classes are used if there are a small number of data points and twenty classes if there are a large number of data points (over 1000 data points). (Note: categories will now be called classes from now on.)
3. $\text{Class width} = \dfrac{\text{range}}{\text{\# classes}}$. Always round up to the next integer (even if the answer is already a whole number go to the next integer). If you don't do this, your last class will not contain your largest data value, and you would have to add another class just for it. If you round up, then your largest data value will fall in the last class, and there are no issues.
4. Create the classes. Each class has limits that determine which values fall in each class. To find the class limits, set the smallest value as the lower class limit for the first class. Then add the class width to the lower class limit to get the next lower class limit. Repeat until you get all the classes. The upper class limit for a class is one less than the lower limit for the next class.
5. In order for the classes to actually touch, then one class needs to start where the previous one ends. This is known as the class boundary. To find the class

boundaries, subtract 0.5 from the lower class limit and add 0.5 to the upper class limit.

6. Sometimes it is useful to find the class midpoint. The process is

$$\text{Midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

7. To figure out the number of data points that fall in each class, go through each data value and see which class boundaries it is between. Utilizing tally marks may be helpful in counting the data values. The frequency for a class is the number of data values that fall in the class.

Note: the above description is for data values that are whole numbers. If you data value has decimal places, then your class width should be rounded up to the nearest value with the same number of decimal places as the original data. In addition, your class boundaries should have one more decimal place than the original data. As an example, if your data have one decimal place, then the class width would have one decimal place, and the class boundaries are formed by adding and subtracting 0.05 from each class limit.

Example #2.2.1: Creating a Frequency Table

Table #2.21 contains the amount of rent paid every month for 24 students from a statistics course. Make a relative frequency distribution using 7 classes.

Table #2.2.1: Data of Monthly Rent

1500	1350	350	1200	850	900
1500	1150	1500	900	1400	1100
1250	600	610	960	890	1325
900	800	2550	495	1200	690

Solution:

1) Find the range:

$$\text{largest value} - \text{smallest value} = 2550 - 350 = 2200$$

2) Pick the number of classes:

The directions say to use 7 classes.

3) Find the class width:

$$\text{width} = \frac{\text{range}}{7} = \frac{2200}{7} \approx 314.286$$

Round up to 315.

Always round up to the next integer even if the width is already an integer.

4) Find the class limits:

Start at the smallest value. This is the lower class limit for the first class. Add the width to get the lower limit of the next class. Keep adding the width to get all the lower limits.

$350 + 315 = 665,\ 665 + 315 = 980,\ 980 + 315 = 1295,\ \ldots$

The upper limit is one less than the next lower limit: so for the first class the upper class limit would be $665 - 1 = 664$.

When you have all 7 classes, make sure the last number, in this case the 2550, is at least as large as the largest value in the data. If not, you made a mistake somewhere.

5) Find the class boundaries:

Subtract 0.5 from the lower class limit to get the class boundaries. Add 0.5 to the upper class limit for the last class's boundary.

$350 - 0.5 = 349.5,\ 665 - 0.5 = 664.5,\ 980 - 0.5 = 979.5,\ 1295 - 0.5 = 1294.5,\ \ldots$

Every value in the data should fall into exactly one of the classes. No data values should fall right on the boundary of two classes.

6) Find the class midpoints:

$$\text{midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

$$\frac{350 + 664}{2} = 507,\ \ \frac{665 + 979}{2} = 822,\ \ldots$$

7) Tally and find the frequency of the data:

Go through the data and put a tally mark in the appropriate class for each piece of data by looking to see which class boundaries the data value is between. Fill in the frequency by changing each of the tallies into a number.

Table #2.2.2: Frequency Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency
350 – 664	349.5 – 664.5	507	𝍸	4
665 – 979	664.5 – 979.5	822	𝍸 III	8
980 – 1294	979.5 – 1294.5	1137	𝍸	5
1295 – 1609	1294.5 – 1609.5	1452	𝍸 I	6
1610 – 1924	1609.5 – 1924.5	1767		0
1925 – 2239	1924.5 – 2239.5	2082		0
2240 – 2554	2239.5 – 2554.5	2397	I	1

Make sure the total of the frequencies is the same as the number of data points.

R command for a frequency distribution:
To create a frequency distribution:
summary(variable) – so you can find out the minimum and maximum.
breaks = seq(min, number above max, by = class width)
breaks – so you can see the breaks that R made.
variable.cut=cut(variable, breaks, right=FALSE) – this will cut up the data into the classes.
variable.freq=table(variable.cut) – this will create the frequency table.
variable.freq – this will display the frequency table.

For the data in Example #2.2.1, the R command would be:

```
rent<-c(1500, 1350, 350, 1200, 850, 900, 1500, 1150, 1500, 900, 1400, 1100,
1250, 600, 610, 960, 890, 1325, 900, 800, 2550, 495, 1200, 690)
summary(rent)
```

Output:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  350.0   837.5  1030.0  1082.0  1331.0  2550.0
```

```
breaks=seq(350, 3000, by = 315)
breaks
```

Output:

```
[1]  350  665  980 1295 1610 1925 2240 2555 2870
```

These are your lower limits of the frequency distribution. You can now write your own table.

```
rent.cut=cut(rent, breaks, right=FALSE)
rent.freq=table(rent.cut)
```

```
rent.freq

Output:
rent.cut
      [350,665)       [665,980)     [980,1.3e+03)  [1.3e+03,1.61e+03)
             4               8               5               6
[1.61e+03,1.92e+03) [1.92e+03,2.24e+03) [2.24e+03,2.56e+03)
[2.56e+03,2.87e+03)
             0               0               1               0
```

It is difficult to determine the basic shape of the distribution by looking at the frequency distribution. It would be easier to look at a graph. The graph of a frequency distribution for quantitative data is called a **frequency histogram** or just histogram for short.

Histogram: a graph of the frequencies on the vertical axis and the class boundaries on the horizontal axis. Rectangles where the height is the frequency and the width is the class width are draw for each class.

Example #2.2.2: Drawing a Histogram

Draw a histogram for the distribution from example #2.2.1.

Solution:

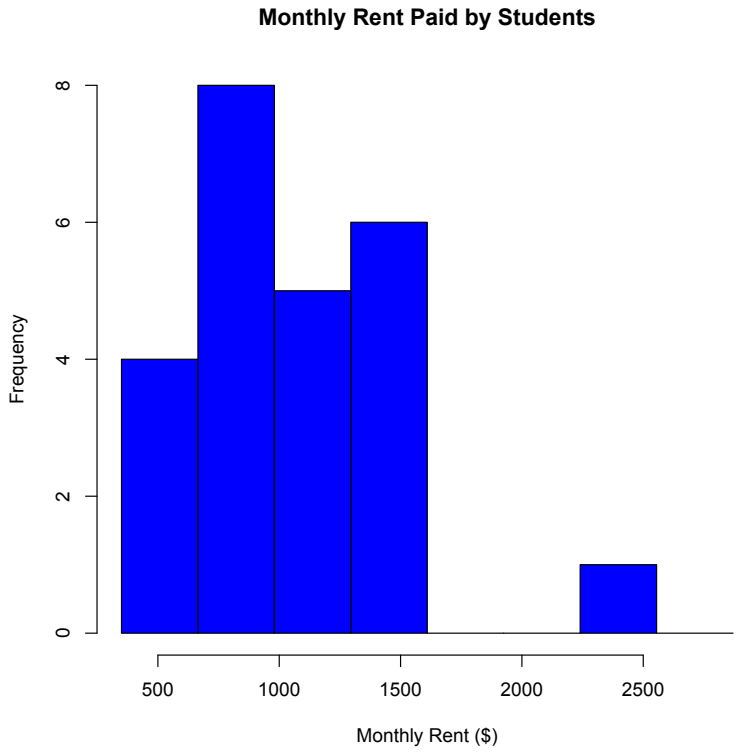
The class boundaries are plotted on the horizontal axis and the frequencies are plotted on the vertical axis. You can plot the midpoints of the classes instead of the class boundaries. Graph #2.2.1 was created using the midpoints because it was easier to do with the software that created the graph. On R, the command is

hist(variable, col="type in what color you want", breaks, main="type the title you want", xlab="type the label you want for the horizontal axis", ylim=c(0, number above maximum frequency) – produces histogram with specified color and using the breaks you made for the frequency distribution.

For this example, the command in R would be (assuming you created a frequency distribution in R as described previously):

hist(rent, col="blue", breaks, right=FALSE, main="Monthly Rent Paid by Students", ylim=c(0,8) xlab="Monthly Rent ($)")

Graph #2.2.1: Histogram for Monthly Rent



If no frequency distribution was created before the histogram, then the command would be:

hist(variable, col="type in what color you want", number of classes, main="type the title you want", xlab="type the label you want for the horizontal axis") – produces histogram with specified color and number of classes (though the number of classes is an estimate and R will create the number of classes near this value).

For this example, the R command without a frequency distribution created first would be:

hist(rent, col="blue", 7, main="Monthly Rent Paid by Students", xlab="Monthly Rent ($)")

Notice the graph has the axes labeled, the tick marks are labeled on each axis, and there is a title.

Reviewing the graph you can see that most of the students pay around $750 per month for rent, with about $1500 being the other common value. You can see from the graph, that most students pay between $600 and $1600 per month for rent. Of course, these values are just estimates from the graph. There is a large

gap between the $1500 class and the highest data value. This seems to say that one student is paying a great deal more than everyone else. This value could be considered an outlier. An **outlier** is a data value that is far from the rest of the values. It may be an unusual value or a mistake. It is a data value that should be investigated. In this case, the student lives in a very expensive part of town, thus the value is not a mistake, and is just very unusual. There are other aspects that can be discussed, but first some other concepts need to be introduced.

Frequencies are helpful, but understanding the relative size each class is to the total is also useful. To find this you can divide the frequency by the total to create a relative frequency. If you have the relative frequencies for all of the classes, then you have a relative frequency distribution.

Relative Frequency Distribution
A variation on a frequency distribution is a relative frequency distribution. Instead of giving the frequencies for each class, the relative frequencies are calculated.

$$\text{Relative frequency} = \frac{\text{frequency}}{\text{\# of data points}}$$

This gives you percentages of data that fall in each class.

Example #2.2.3: Creating a Relative Frequency Table

Find the relative frequency for the grade data.

Solution:

From example #2.2.1, the frequency distribution is reproduced in table #2.2.2.

Table #2.2.2: Frequency Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Frequency
350 – 664	349.5 – 664.5	507	4
665 – 979	664.5 – 979.5	822	8
980 – 1294	979.5 – 1294.5	1137	5
1295 – 1609	1294.5 – 1609.5	1452	6
1610 – 1924	1609.5 – 1924.5	1767	0
1925 – 2239	1924.5 – 2239.5	2082	0
2240 – 2554	2239.5 – 2554.5	2397	1

Divide each frequency by the number of data points.

$\frac{4}{24} = 0.17,\ \frac{8}{24} = 0.33,\ \frac{5}{24} = 0.21,\ \ldots$

Table #2.2.3: Relative Frequency Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Frequency	Relative Frequency
350 – 664	349.5 – 664.5	507	4	0.17
665 – 979	664.5 – 979.5	822	8	0.33
980 – 1294	979.5 – 1294.5	1137	5	0.21
1295 – 1609	1294.5 – 1609.5	1452	6	0.25
1610 – 1924	1609.5 – 1924.5	1767	0	0
1925 – 2239	1924.5 – 2239.5	2082	0	0
2240 – 2554	2239.5 – 2554.5	2397	1	0.04
Total			24	1

The relative frequencies should add up to 1 or 100%. (This might be off a little due to rounding errors.)

The graph of the relative frequency is known as a relative frequency histogram. It looks identical to the frequency histogram, but the vertical axis is relative frequency instead of just frequencies.

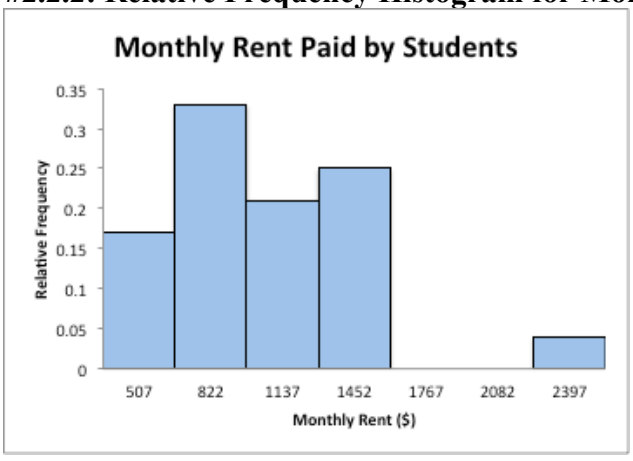
Example #2.2.4: Drawing a Relative Frequency Histogram

Draw a relative frequency histogram for the grade distribution from example #2.2.1.

Solution:

The class boundaries are plotted on the horizontal axis and the relative frequencies are plotted on the vertical axis. (This is not easy to do in R, so use another technology to graph a relative frequency histogram.)

Graph #2.2.2: Relative Frequency Histogram for Monthly Rent



Notice the shape is the same as the frequency distribution.

Another useful piece of information is how many data points fall below a particular class boundary. As an example, a teacher may want to know how many students received below an 80%, a doctor may want to know how many adults have cholesterol below 160, or a manager may want to know how many stores gross less than $2000 per day. This is known as a **cumulative frequency**. If you want to know what percent of the data falls below a certain class boundary, then this would be a **cumulative relative frequency**. For cumulative frequencies you are finding how many data values fall below the upper class limit.

To create a **cumulative frequency distribution**, count the number of data points that are below the upper class boundary, starting with the first class and working up to the top class. The last upper class boundary should have all of the data points below it. Also include the number of data points below the lowest class boundary, which is zero.

Example #2.2.5: Creating a Cumulative Frequency Distribution

Create a cumulative frequency distribution for the data in example #2.2.1.

Solution:

The frequency distribution for the data is in table #2.2.2.

Table #2.2.2: Frequency Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Frequency
350 – 664	349.5 – 664.5	507	4
665 – 979	664.5 – 979.5	822	8
980 – 1294	979.5 – 1294.5	1137	5
1295 – 1609	1294.5 – 1609.5	1452	6
1610 – 1924	1609.5 – 1924.5	1767	0
1925 – 2239	1924.5 – 2239.5	2082	0
2240 – 2554	2239.5 – 2554.5	2397	1

Now ask yourself how many data points fall below each class boundary. Below 349.5, there are 0 data points. Below 664.5 there are 4 data points, below 979.5, there are 4 + 8 = 12 data points, below 1294.5 there are 4 + 8 + 5 = 17 data points, and continue this process until you reach the upper class boundary. This is summarized in Table #2.2.4.

To produce cumulative frequencies in R, you need to have performed the commands for the frequency distribution. Once you have complete that, then use variable.cumfreq=cumsum(variable.freq) – creates the cumulative frequencies for the variable

cumfreq0=c(0,variable.cumfreq) – creates a cumulative frequency table for the variable.
cumfreq0 – displays the cumulative frequency table.

For this example the command would be:
rent.cumfreq=cumsum(rent.freq)
cumfreq0=c(0,rent.cumfreq)
cumfreq0

Output:

```
                [350,665)         [665,980)      [980,1.3e+03)
         0              4              12              17
 [1.3e+03,1.61e+03) [1.61e+03,1.92e+03) [1.92e+03,2.24e+03)
[2.24e+03,2.56e+03)
        23             23              23              24
[2.56e+03,2.87e+03)
        24
```

Now type this into a table. See Table #2.2.4.

Table #2.2.4: Cumulative Distribution for Monthly Rent

Class Limits	Class Boundaries	Class Midpoint	Frequency	Cumulative Frequency
350 – 664	349.5 – 664.5	507	4	4
665 – 979	664.5 – 979.5	822	8	12
980 – 1294	979.5 – 1294.5	1137	5	17
1295 – 1609	1294.5 – 1609.5	1452	6	23
1610 – 1924	1609.5 – 1924.5	1767	0	23
1925 – 2239	1924.5 – 2239.5	2082	0	23
2240 – 2554	2239.5 – 2554.5	2397	1	24

Again, it is hard to look at the data the way it is. A graph would be useful. The graph for cumulative frequency is called an **ogive** (o-jive). To create an ogive, first create a scale on both the horizontal and vertical axes that will fit the data. Then plot the points of the class upper class boundary versus the cumulative frequency. Make sure you include the point with the lowest class boundary and the 0 cumulative frequency. Then just connect the dots.

Example #2.2.6: Drawing an Ogive

Draw an ogive for the data in example #2.2.1.

Solution:

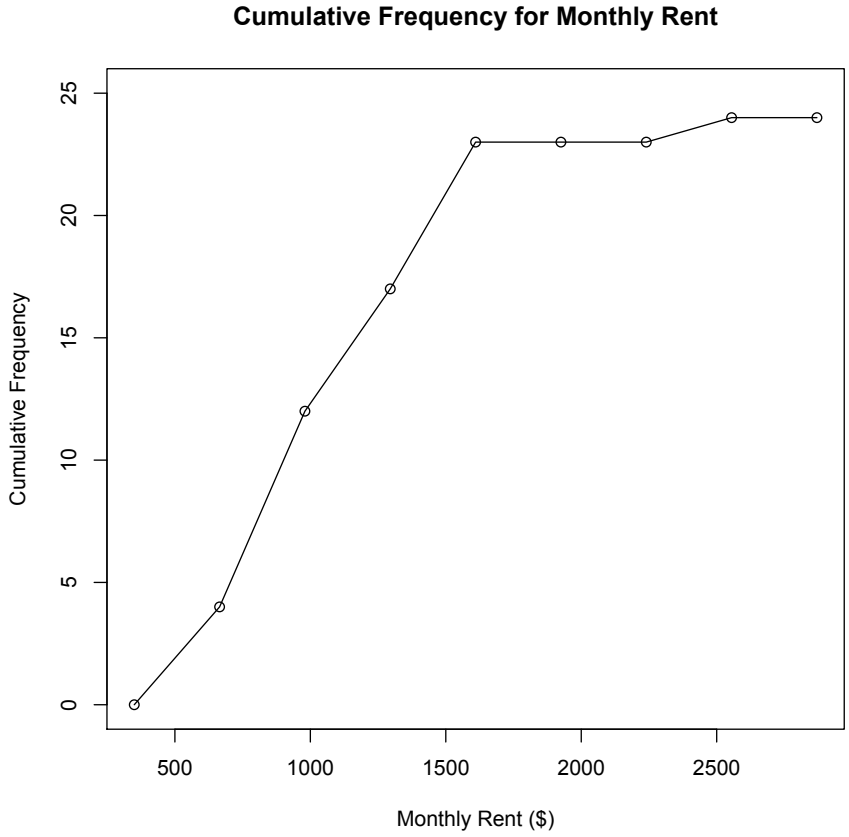
In R, the commands would be:

plot(breaks,cumfreq0, main="title you want to use", xlab="label you want to use", ylab="label you want to use", ylim=c(0, number above maximum cumulative frequency) – plots the ogive
lines(breaks,cumfreq0) – connects the dots on the ogive

For this example, the commands would be:

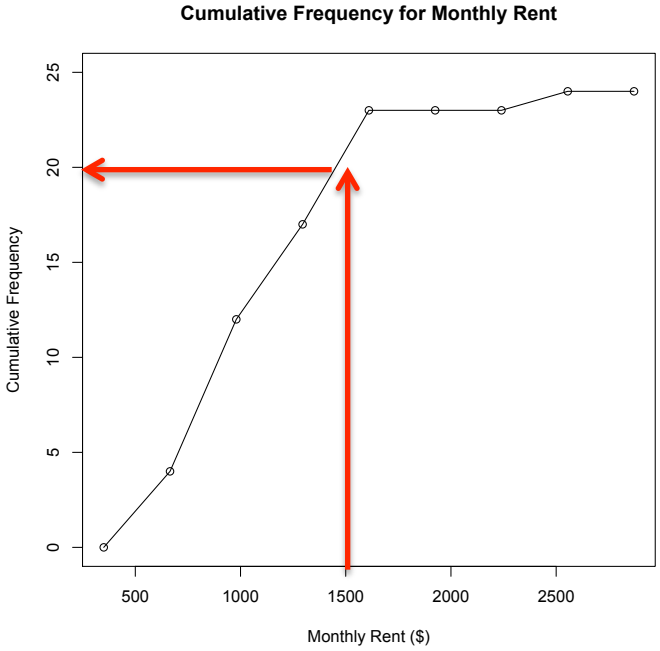
Plot(breaks,cumfreq0, main="Cumulative Frequency for Monthly Rent", xlab="Monthly Rent ($)", ylab="Cumulative Frequency", ylim=c(0,25))
lines(breaks,cumfreq0)

Graph #2.2.3: Ogive for Monthly Rent



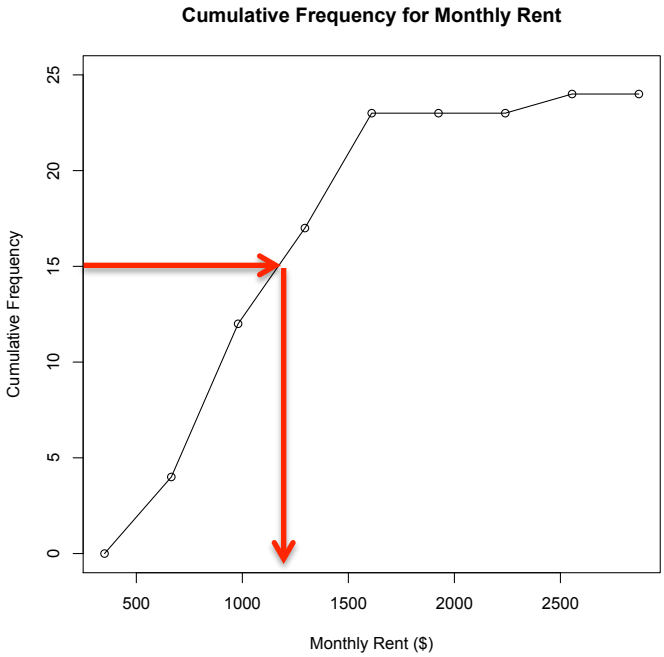
The usefulness of a ogive is to allow the reader to find out how many students pay less than a certain value, and also what amount of monthly rent is paid by a certain number of students. As an example, suppose you want to know how many students pay less than $1500 a month in rent, then you can go up from the $1500 until you hit the graph and then you go over to the cumulative frequency axes to see what value corresponds to this value. It appears that around 20 students pay less than $1500. (See graph #2.2.4.)

Graph #2.2.4: Ogive for Monthly Rent with Example



Also, if you want to know the amount that 15 students pay less than, then you start at 15 on the vertical axis and then go over to the graph and down to the horizontal axis where the line intersects the graph. You can see that 15 students pay less than about $1200 a month. (See graph #2.2.5.)

Graph #2.2.5: Ogive for Monthly Rent with Example



If you graph the cumulative relative frequency then you can find out what percentage is below a certain number instead of just the number of people below a certain value.

Shapes of the distribution:

When you look at a distribution, look at the basic shape. There are some basic shapes that are seen in histograms. Realize though that some distributions have no shape. The common shapes are symmetric, skewed, and uniform. Another interest is how many peaks a graph may have. This is known as modal.

Symmetric means that you can fold the graph in half down the middle and the two sides will line up. You can think of the two sides as being mirror images of each other. Skewed means one "tail" of the graph is longer than the other. The graph is skewed in the direction of the longer tail (backwards from what you would expect). A uniform graph has all the bars the same height.

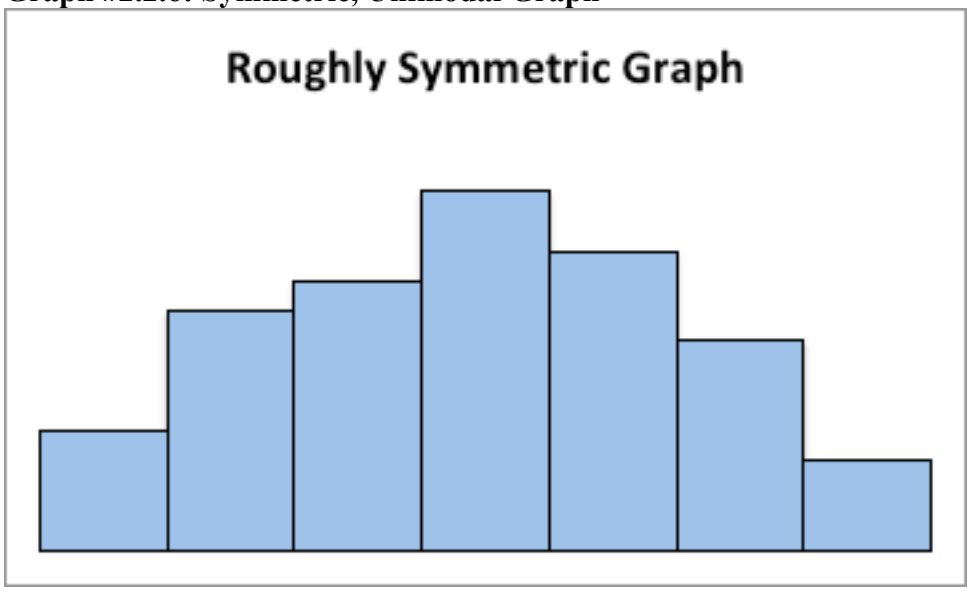
Modal refers to the number of peaks. Unimodal has one peak and bimodal has two peaks. Usually if a graph has more than two peaks, the modal information is not longer of interest.

Other important features to consider are gaps between bars, a repetitive pattern, how spread out is the data, and where the center of the graph is.

Examples of graphs:

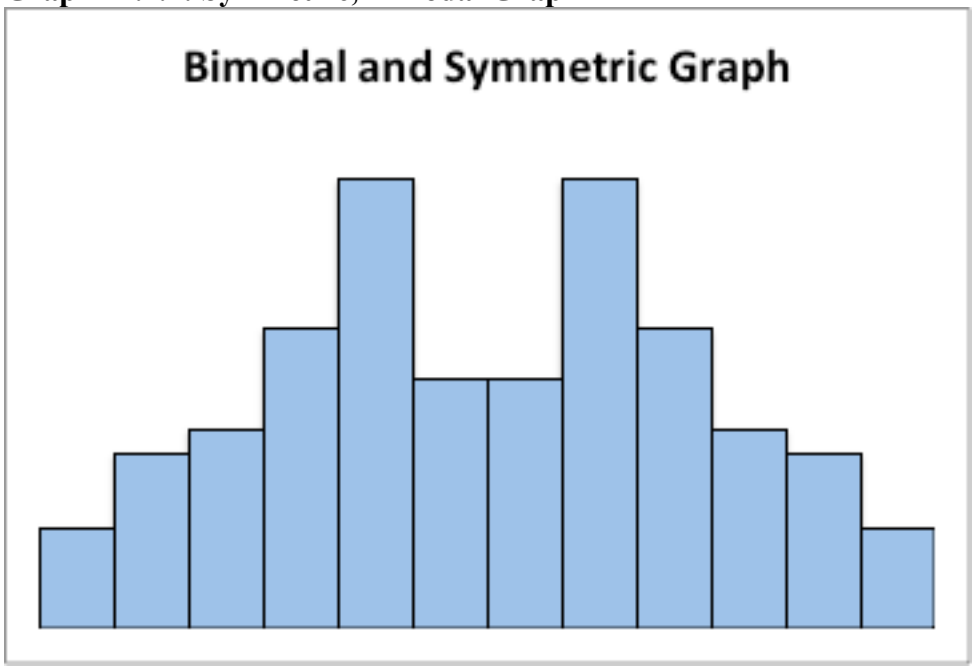
This graph is roughly symmetric and unimodal:

Graph #2.2.6: Symmetric, Unimodal Graph



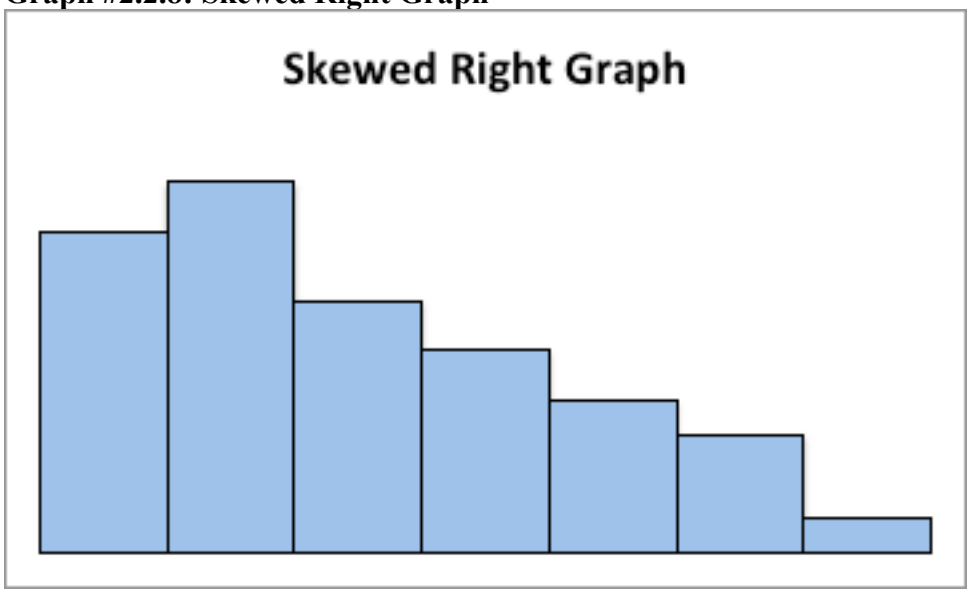
This graph is symmetric and bimodal:

Graph #2.2.7: Symmetric, Bimodal Graph



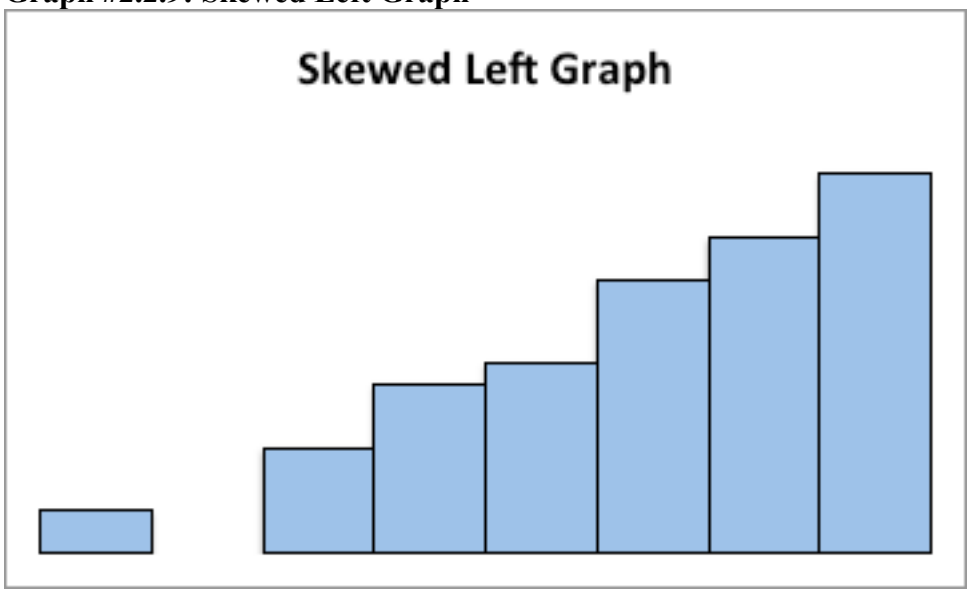
This graph is skewed to the right:

Graph #2.2.8: Skewed Right Graph



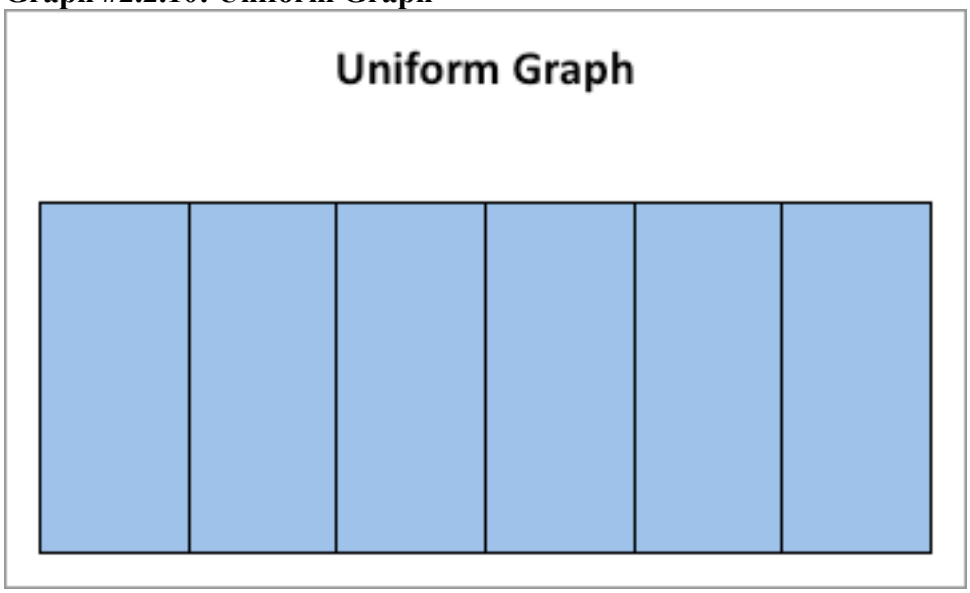
This graph is skewed to the left and has a gap:

Graph #2.2.9: Skewed Left Graph



This graph is uniform since all the bars are the same height:

Graph #2.2.10: Uniform Graph



Example #2.2.7: Creating a Frequency Distribution, Histogram, and Ogive

The following data represents the percent change in tuition levels at public, four-year colleges (inflation adjusted) from 2008 to 2013 (Weissmann, 2013). Create a frequency distribution, histogram, and ogive for the data.

Table #2.2.5: Data of Tuition Levels at Public, Four-Year Colleges

19.5%	40.8%	57.0%	15.1%	17.4%	5.2%	13.0%	15.6%
51.5%	15.6%	14.5%	22.4%	19.5%	31.3%	21.7%	27.0%
13.1%	26.8%	24.3%	38.0%	21.1%	9.3%	46.7%	14.5%
78.4%	67.3%	21.1%	22.4%	5.3%	17.3%	17.5%	36.6%
72.0%	63.2%	15.1%	2.2%	17.5%	36.7%	2.8%	16.2%
20.5%	17.8%	30.1%	63.6%	17.8%	23.2%	25.3%	21.4%
28.5%	9.4%						

Solution:

1) Find the range:

$$\text{largest value} - \text{smallest value} = 78.4\% - 2.2\% = 76.2\%$$

2) Pick the number of classes:

Since there are 50 data points, then around 6 to 8 classes should be used. Let's use 8.

3) Find the class width:

$$\text{width} = \frac{\text{range}}{8} = \frac{76.2\%}{8} \approx 9.525\%$$

Since the data has one decimal place, then the class width should round to one decimal place. Make sure you round up.

width = 9.6%

4) Find the class limits:

$$2.2\% + 9.6\% = 11.8\%,\ 11.8\% + 9.6\% = 21.4\%,\ 21.4\% + 9.6\% = 31.0\%,\ldots$$

5) Find the class boundaries:

Since the data has one decimal place, the class boundaries should have two decimal places, so subtract 0.05 from the lower class limit to get the class boundaries. Add 0.05 to the upper class limit for the last class's boundary.
$2.2 - 0.05 = 2.15\%,\ 11.8 - 0.05 = 11.75\%,\ 21.4 - 0.05 = 21.35\%,\ \ldots$
Every value in the data should fall into exactly one of the classes. No data values should fall right on the boundary of two classes.

6) Find the class midpoints:

$$\text{midpoint} = \frac{\text{lower limt } + \text{ upper limit}}{2}$$

$$\frac{2.2+11.7}{2} = 6.95\%,\ \ \frac{11.8+21.3}{2} = 16.55\%,\ \ldots$$

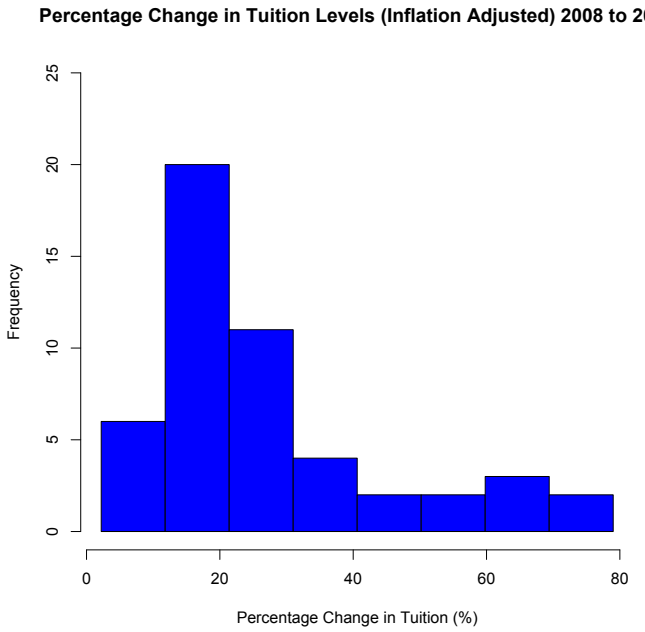
7) Tally and find the frequency of the data:

Table #2.2.6: Frequency Distribution for Tuition Levels at Public, Four-Year Colleges

Class Limits	Class Boundaries	Class Midpoint	Tally	Frequency	Relative Frequency	Cumulative Frequency
2.2 – 11.7	2.15 – 11.75	6.95	𝍸 I	6	0.12	6
11.8 – 21.3	11.75 – 21.35	16.55	𝍸 𝍸 𝍸 𝍸	20	0.40	26
21.4 – 30.9	21.35 – 30.95	26.15	𝍸 𝍸 I	11	0.22	37
31.0 – 40.5	30.95 – 40.55	35.75	IIII	4	0.08	41
40.6 – 50.1	40.55 – 50.15	45.35	II	2	0.04	43
50.2 – 59.7	50.15 – 59.75	54.95	II	2	0.04	45
59.8 – 69.3	59.75 – 69.35	64.55	III	3	0.06	48
69.4 – 78.9	69.35 – 78.95	74.15	II	2	0.04	50

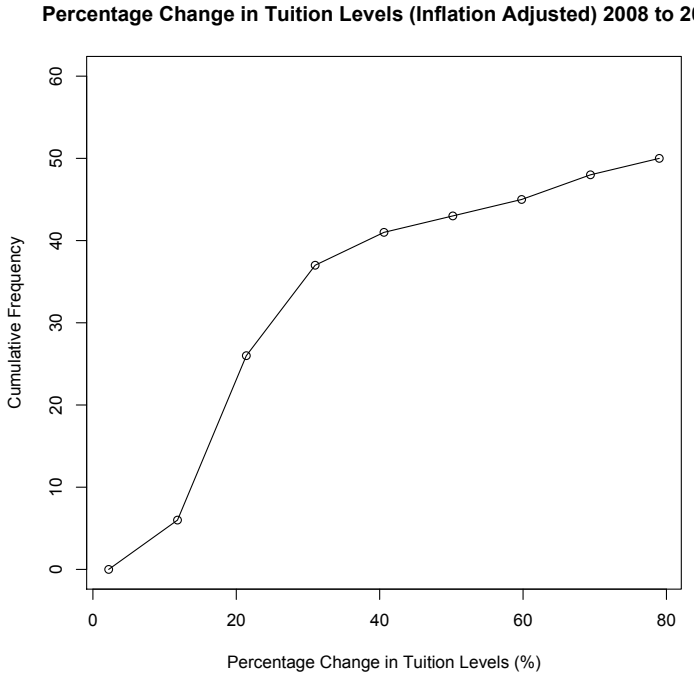
Make sure the total of the frequencies is the same as the number of data points.

Graph #2.2.11: Histogram for Tuition Levels at Public, Four-Year Colleges



This graph is skewed right, with no gaps. This says that most percent increases in tuition were around 16.55%, with very few states having a percent increase greater than 45.35%.

Graph #2.2.11: Ogive for Tuition Levels at Public, Four-Year Colleges



Looking at the ogive, you can see that 30 states had a percent change in tuition levels of about 25% or less.

There are occasions where the class limits in the frequency distribution are predetermined. Example #2.2.8 demonstrates this situation.

Example #2.2.8: Creating a Frequency Distribution and Histogram

The following are the percentage grades of 25 students from a statistics course. Make a frequency distribution and histogram.

Table #2.2.7: Data of Test Grades

62	87	81	69	87	62	45	95	76	76
62	71	65	67	72	80	40	77	87	58
84	73	93	64	89					

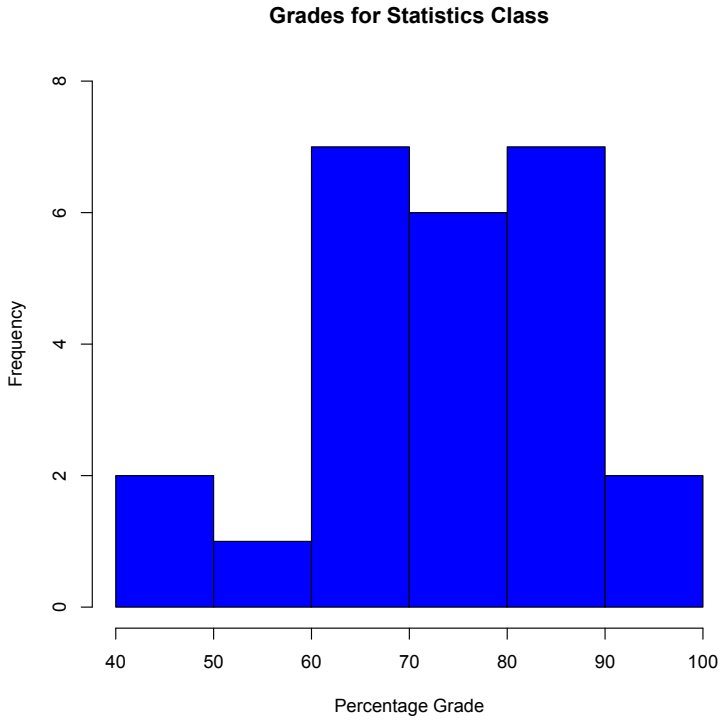
Solution:

Since this data is percent grades, it makes more sense to make the classes in multiples of 10, since grades are usually 90 to 100%, 80 to 90%, and so forth. It is easier to not use the class boundaries, but instead use the class limits and think of the upper class limit being up to but not including the next classes lower limit. As an example the class 80 – 90 means a grade of 80% up to but not including a 90%. A student with an 89.9% would be in the 80-90 class.

Table #2.2.8: Frequency Distribution for Test Grades

Class Limit	Class Midpoint	Tally	Frequency
40 – 50	45	||	2
50 – 60	55			1
60 – 70	65	𝍸 ||	7
70 – 80	75	𝍸		6
80 – 90	85	𝍸 ||	7
90 – 100	95	||	2

Graph #2.2.12: Histogram for Test Grades



It appears that most of the students had between 60 to 90%. This graph looks somewhat symmetric and also bimodal. The same number of students earned between 60 to 70% and 80 to 90%.

There are other types of graphs for quantitative data. They will be explored in the next section.

Section 2.2: Homework

1.) The median incomes of males in each state of the United States, including the District of Columbia and Puerto Rico, are given in table #2.2.9 ("Median income of," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 7 classes.

Table #2.2.9: Data of Median Income for Males

$42,951	$52,379	$42,544	$37,488	$49,281	$50,987	$60,705
$50,411	$66,760	$40,951	$43,902	$45,494	$41,528	$50,746
$45,183	$43,624	$43,993	$41,612	$46,313	$43,944	$56,708
$60,264	$50,053	$50,580	$40,202	$43,146	$41,635	$42,182
$41,803	$53,033	$60,568	$41,037	$50,388	$41,950	$44,660
$46,176	$41,420	$45,976	$47,956	$22,529	$48,842	$41,464
$40,285	$41,309	$43,160	$47,573	$44,057	$52,805	$53,046
$42,125	$46,214	$51,630				

2.) The median incomes of females in each state of the United States, including the District of Columbia and Puerto Rico, are given in table #2.2.10 ("Median income of," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 7 classes.

Table #2.2.10: Data of Median Income for Females

$31,862	$40,550	$36,048	$30,752	$41,817	$40,236	$47,476	$40,500
$60,332	$33,823	$35,438	$37,242	$31,238	$39,150	$34,023	$33,745
$33,269	$32,684	$31,844	$34,599	$48,748	$46,185	$36,931	$40,416
$29,548	$33,865	$31,067	$33,424	$35,484	$41,021	$47,155	$32,316
$42,113	$33,459	$32,462	$35,746	$31,274	$36,027	$37,089	$22,117
$41,412	$31,330	$31,329	$33,184	$35,301	$32,843	$38,177	$40,969
$40,993	$29,688	$35,890	$34,381				

3.) The density of people per square kilometer for African countries is in table #2.2.11 ("Density of people," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 8 classes.

Table #2.2.11: Data of Density of People per Square Kilometer

15	16	81	3	62	367	42	123
8	9	337	12	29	70	39	83
26	51	79	6	157	105	42	45
72	72	37	4	36	134	12	3
630	563	72	29	3	13	176	341
415	187	65	194	75	16	41	18
69	49	103	65	143	2	18	31

4.) The Affordable Care Act created a market place for individuals to purchase health care plans. In 2014, the premiums for a 27 year old for the bronze level health insurance are given in table #2.2.12 ("Health insurance marketplace," 2013). Create a frequency distribution, relative frequency distribution, and cumulative frequency distribution using 5 classes.

Table #2.2.12: Data of Health Insurance Premiums

$114	$119	$121	$125	$132	$139
$139	$141	$143	$145	$151	$153
$156	$159	$162	$163	$165	$166
$170	$170	$176	$177	$181	$185
$185	$186	$186	$189	$190	$192
$196	$203	$204	$219	$254	$286

5.) Create a histogram and relative frequency histogram for the data in table #2.2.9. Describe the shape and any findings you can from the graph.

6.) Create a histogram and relative frequency histogram for the data in table #2.2.10. Describe the shape and any findings you can from the graph.

7.) Create a histogram and relative frequency histogram for the data in table #2.2.11. Describe the shape and any findings you can from the graph.

8.) Create a histogram and relative frequency histogram for the data in table #2.2.12. Describe the shape and any findings you can from the graph.

9.) Create an ogive for the data in table #2.2.9. Describe any findings you can from the graph.

10.) Create an ogive for the data in table #2.2.10. Describe any findings you can from the graph.

11.) Create an ogive for the data in table #2.2.11. Describe any findings you can from the graph.

12.) Create an ogive for the data in table #2.2.12. Describe any findings you can from the graph.

13.) Students in a statistics class took their first test. The following are the scores they earned. Create a frequency distribution and histogram for the data using class limits that make sense for grade data. Describe the shape of the distribution.

Table #2.2.13: Data of Test 1 Grades

80	79	89	74	73	67	79
93	70	70	76	88	83	73
81	79	80	85	79	80	79
58	93	94	74			

14.) Students in a statistics class took their first test. The following are the scores they earned. Create a frequency distribution and histogram for the data using class limits that make sense for grade data. Describe the shape of the distribution. Compare to the graph in question 13.

Table #2.2.14: Data of Test 1 Grades

67	67	76	47	85	70
87	76	80	72	84	98
84	64	65	82	81	81
88	74	87	83		

Section 2.3: Other Graphical Representations of Data

There are many other types of graphs. Some of the more common ones are the frequency polygon, the dot plot, the stem plot, scatter plot, and a time-series plot. There are also many different graphs that have emerged lately for qualitative data. Many are found in publications and websites. The following is a description of the stem plot, the scatter plot, and the time-series plot.

Stem Plots

Stem plots are a quick and easy way to look at small samples of numerical data. You can look for any patterns or any strange data values. It is easy to compare two samples using stem plots.

The first step is to divide each number into 2 parts, the stem (such as the leftmost digit) and the leaf (such as the rightmost digit). There are no set rules, you just have to look at the data and see what makes sense.

Example #2.3.1: Stem Plot for Grade Distribution

The following are the percentage grades of 25 students from a statistics course. Draw a stem plot of the data.

Table #2.3.1: Data of Test Grades

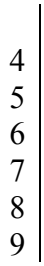
62	87	81	69	87	62	45	95	76	76
62	71	65	67	72	80	40	77	87	58
84	73	93	64	89					

Solution:

Divide each number so that the tens digit is the stem and the ones digit is the leaf. 62 becomes 6|2.

Make a vertical chart with the stems on the left of a vertical bar. Be sure to fill in any missing stems. In other words, the stems should have equal spacing (for example, count by ones or count by tens). The graph #2.3.1 shows the stems for this example.

Graph #2.3.1: Stem plot for Test Grades Step 1



Now go through the list of data and add the leaves. Put each leaf next to its corresponding stem. Don't worry about order yet just get all the leaves down.

When the data value 62 is placed on the plot it looks like the plot in graph #2.3.2.

Graph #2.3.2: Stem plot for Test Grades Step 2

4	
5	
6	2
7	
8	
9	

When the data value 87 is placed on the plot it looks like the plot in graph #2.3.3.

Graph #2.3.3: Stem plot for Test Grades Step 3

4	
5	
6	2
7	
8	7
9	

Filling in the rest of the leaves to obtain the plot in graph #2.3.4.

Graph #2.3.4: Stem plot for Test Grades Step 4

4	5 0
5	8
6	2 9 2 2 5 7 4
7	6 6 1 2 7 3
8	7 1 7 0 7 4 9
9	5 3

Now you have to add labels and make the graph look pretty. You need to add a label and sort the leaves into increasing order. You also need to tell people what the stems and leaves mean by inserting a legend. **Be careful to line the leaves up in columns.** You need to be able to compare the lengths of the rows when you interpret the graph. The final stem plot for the test grade data is in graph #2.3.5.

Graph #2.3.5: Stem plot for Test Grades

Test Scores

4	0 = 40%
4	0 5
5	8
6	2 2 2 4 5 7 9
7	1 2 3 6 6 7
8	0 1 4 7 7 7 9
9	3 5

Now you can interpret the stem-and-leaf display. The data is bimodal and somewhat symmetric. There are no gaps in the data. The center of the distribution is around 70.

You can create a stem and leaf plot on R. the command is:
stem(variable) – creates a stem and leaf plot, if you do not get a stem plot that shows all of the stems then use scale = a number. Adjust the number until you see all of the stems. So you would have stem(variable, scale = a number)

For Example #2.3.1, the command would be

```
grades<-c(62, 87, 81, 69, 87, 62, 45, 95, 76, 76, 62, 71, 65, 67, 72, 80, 40, 77, 87,
58, 84, 73, 93, 64, 89)
stem(grades, scale = 2)
```

Output:

The decimal point is 1 digit(s) to the right of the |

```
4 | 05
5 | 8
6 | 2224579
7 | 123667
8 | 0147779
9 | 35
```

Now just put a title on the stem plot

Scatter Plot

Sometimes you have two different variables and you want to see if they are related in any way. A scatter plot helps you to see what the relationship would look like. A scatter plot is just a plotting of the ordered pairs.

Example #2.3.2: Scatter Plot

Is there any relationship between elevation and high temperature on a given day? The following data are the high temperatures at various cities on a single day and the elevation of the city.

Table #2.3.2: Data of Temperature versus Elevation

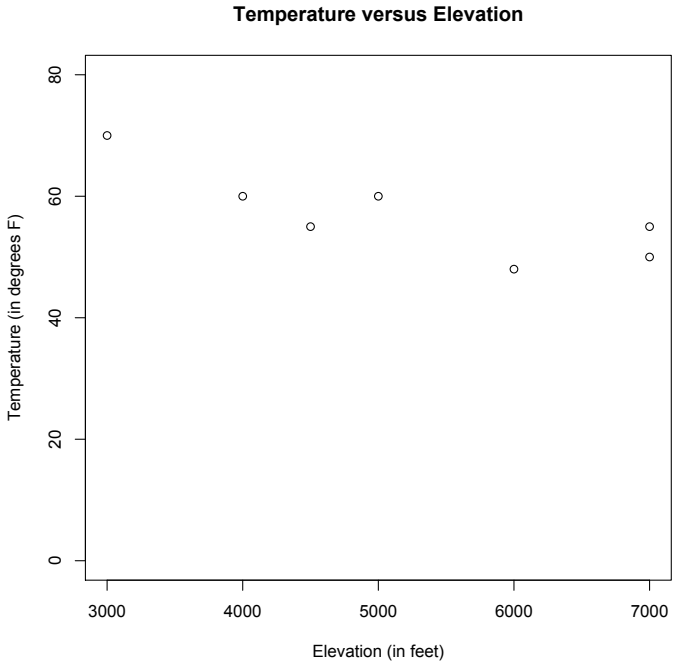
Elevation (in feet)	7000	4000	6000	3000	7000	4500	5000
Temperature (°F)	50	60	48	70	55	55	60

Solution:
Preliminary: State the random variables
Let x = altitude
y = high temperature
Now plot the x values on the horizontal axis, and the y values on the vertical axis. Then set up a scale that fits the data on each axes. Once that is done, then just plot the x and y values as an ordered pair. In R, the command is:

independent variable<-c(type in data with commas in between values)
dependent variable<-c(type in data with commas in between values)
plot(independent variable, dependent variable, main="type in a title you want", xlab="type in a label for the horizontal axis", ylab="type in a label for the vertical axis", ylim=c(0, number above maximum y value)

For this example, that would be:
elevation<-c(7000, 4000, 6000, 3000, 7000, 4500, 5000)
temperature<-c(50, 60, 48, 70, 55, 55, 60)
plot(elevation, temperature, main="Temperature versus Elevation",
xlab="Elevation (in feet)", ylab="Temperature (in degrees F)", ylim=c(0, 80))

Graph #2.3.6: Scatter Plot of Temperature versus Elevation



Looking at the graph, it appears that there is a linear relationship between temperature and elevation. It also appears to be a negative relationship, thus as elevation increases, the temperature decreases.

Time-Series

A time-series plot is a graph showing the data measurements in chronological order, the data being quantitative data. For example, a time-series plot is used to show profits over the last 5 years. To create a time-series plot, the time always goes on the horizontal axis, and the other variable goes on the vertical axis. Then plot the ordered pairs and connect the dots. The purpose of a time-series graph is to look for trends over time. Caution, you must realize that the trend may not continue. Just because you see an increase, doesn't mean the increase will continue forever. As an example, prior to 2007, many people noticed that housing prices were increasing. The belief at the time was that housing prices would continue to increase. However, the housing bubble burst in 2007, and many houses lost value, and haven't recovered.

Example #2.3.3: Time-Series Plot

The following table tracks the weight of a dieter, where the time in months is measuring how long since the person started the diet.

Table #2.3.3: Data of Weights versus Time

Time (months)	0	1	2	3	4	5
Weight (pounds)	200	195	192	193	190	187

Make a time-series plot of this data

Solution:

In R, the command would be:

variable1<-c(type in data with commas in between values, this should be the time variable)

variable2<-c(type in data with commas in between values)

plot(variable1, variable2, ylim=c(0,number over max), main="type in a title you want", xlab="type in a label for the horizontal axis", ylab="type in a label for the vertical axis")

lines(variable1, variable2) – connects the dots

For this example:

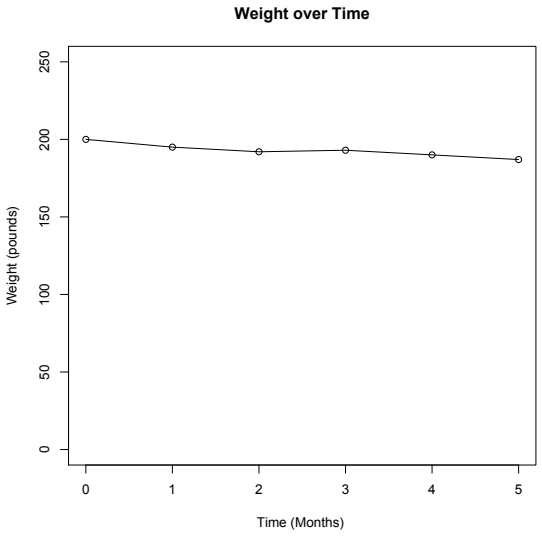
time<-c(0, 1, 2, 3, 4, 5)

weight<-c(200, 195, 192, 193, 190, 187)

plot(time, weight, ylim=c(0,250), main="Weight over Time", xlab="Time (Months) ", ylab="Weight (pounds)")

lines(time, weight)

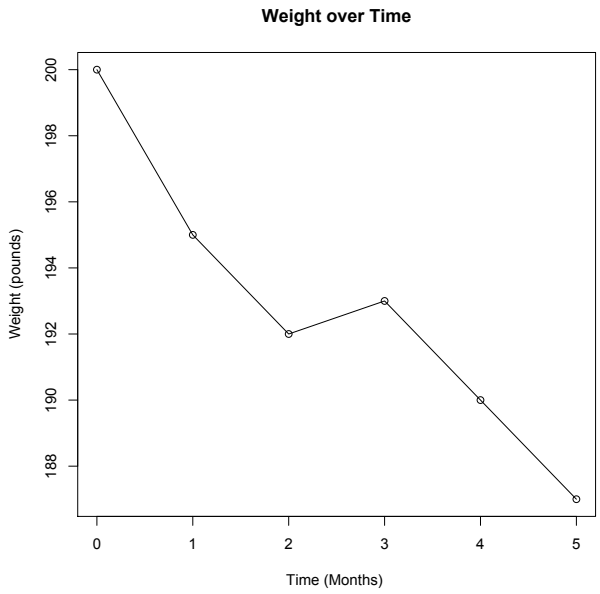
Graph #2.3.7: Time-Series Graph of Weight versus Time



Notice, that over the 5 months, the weight appears to be decreasing. Though it doesn't look like there is a large decrease.

Be careful when making a graph. If you don't start the vertical axis at 0, then the change can look much more dramatic than it really is. As an example, graph #2.3.8 shows the graph #2.3.7 with a different scaling on the vertical axis. Notice the decrease in weight looks much larger than it really is.

Graph #2.3.8: Example of a Poor Graph



Section 2.3: Homework

1.) Students in a statistics class took their first test. The data in table #2.3.4 are the scores they earned. Create a stem plot.

Table #2.3.4: Data of Test 1 Grades

80	79	89	74	73	67	79
93	70	70	76	88	83	73
81	79	80	85	79	80	79
58	93	94	74			

2.) Students in a statistics class took their first test. The data in table #2.3.5 are the scores they earned. Create a stem plot. Compare to the graph in question 1.

Table #2.3.5: Data of Test 1 Grades

67	67	76	47	85	70
87	76	80	72	84	98
84	64	65	82	81	81
88	74	87	83		

3.) When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of one of their metacarpal bone (in cm) were collected and are in table #2.4.6 ("Prediction of height," 2013). Create a scatter plot and state if there is a relationship between the height of a person and the length of their metacarpal.

Table #2.3.6: Data of Metacarpal versus Height

Length of Metacarpal	Height of Person
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

4.) Table #2.3.7 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013). Create a scatter plot and state if there is a relationship between the value of the house and the annual rental income.

Table #2.3.7: Data of House Value versus Rental

Value	Rental	Value	Rental	Value	Rental	Value	Rental
81000	6656	77000	4576	75000	7280	67500	6864
95000	7904	94000	8736	90000	6240	85000	7072
121000	12064	115000	7904	110000	7072	104000	7904
135000	8320	130000	9776	126000	6240	125000	7904
145000	8320	140000	9568	140000	9152	135000	7488
165000	13312	165000	8528	155000	7488	148000	8320
178000	11856	174000	10400	170000	9568	170000	12688
200000	12272	200000	10608	194000	11232	190000	8320
214000	8528	208000	10400	200000	10400	200000	8320
240000	10192	240000	12064	240000	11648	225000	12480
289000	11648	270000	12896	262000	10192	244500	11232
325000	12480	310000	12480	303000	12272	300000	12480

5.) The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in table #2.3.8. Create a scatter plot of the data and state if there appears to be a relationship between life expectancy and the number of births per woman.

Table #2.3.8: Data of Life Expectancy versus Fertility Rate

Life Expectancy	Fertility Rate	Life Expectancy	Fertility Rate
77.2	1.7	72.3	3.9
55.4	5.8	76.0	1.5
69.9	2.2	66.0	4.2
76.4	2.1	55.9	5.2
75.0	1.8	54.4	6.8
78.2	2.0	62.9	4.7
73.0	2.6	78.3	2.1
70.8	2.8	72.1	2.9
82.6	1.4	80.7	1.4
68.9	2.6	74.2	2.5
81.0	1.5	73.3	1.5
54.2	6.9	67.1	2.4

6.) The World Bank collected data on the percentage of gross domestic product (GDP) that a country spends on health expenditures ("Health expenditure," 2013) and the percentage of woman receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 is in table #2.3.9. Create a scatter plot of the data and state if there appears to be a relationship between percentage spent on health expenditure and the percentage of woman receiving prenatal care.

Table #2.3.9: Data of Prenatal Care versus Health Expenditure

Prenatal Care (%)	Health Expenditure (% of GDP)
47.9	9.6
54.6	3.7
93.7	5.2
84.7	5.2
100.0	10.0
42.5	4.7
96.4	4.8
77.1	6.0
58.3	5.4
95.4	4.8
78.0	4.1
93.3	6.0
93.3	9.5
93.7	6.8
89.8	6.1

7.) The Australian Institute of Criminology gathered data on the number of deaths (per 100,000 people) due to firearms during the period 1983 to 1997 ("Deaths from firearms," 2013). The data is in table #2.3.10. Create a time-series plot of the data and state any findings you can from the graph.

Table #2.3.10: Data of Year versus Number of Deaths due to Firearms

Year	1983	1984	1985	1986	1987	1988	1989	1990
Rate	4.31	4.42	4.52	4.35	4.39	4.21	3.40	3.61
Year	1991	1992	1993	1994	1995	1996	1997	
Rate	3.67	3.61	2.98	2.95	2.72	2.95	2.3	

8.) The economic crisis of 2008 affected many countries, though some more than others. Some people in Australia have claimed that Australia wasn't hurt that badly from the crisis. The bank assets (in billions of Australia dollars (AUD)) of the Reserve Bank of Australia (RBA) for the time period of March 2007 through March 2013 are contained in table #2.3.11 ("B1 assets of," 2013). Create a time-series plot and interpret any findings.

Table #2.3.11: Data of Date versus RBA Assets

Date	Assets in billions of AUD
Mar-2006	96.9
Jun-2006	107.4
Sep-2006	107.2
Dec-2006	116.2
Mar-2007	123.7
Jun-2007	134.0
Sep-2007	123.0
Dec-2007	93.2
Mar-2008	93.7
Jun-2008	105.6
Sep-2008	101.5
Dec-2008	158.8
Mar-2009	118.7
Jun-2009	111.9
Sep-2009	87.0
Dec-2009	86.1
Mar-2010	83.4
Jun-2010	85.7
Sep-2010	74.8
Dec-2010	76.0
Mar-2011	75.7
Jun-2011	75.9
Sep-2011	75.2
Dec-2011	87.9
Mar-2012	91.0
Jun-2012	90.1
Sep-2012	83.9
Dec-2012	95.8
Mar-2013	90.5

9.) The consumer price index (CPI) is a measure used by the U.S. government to describe the cost of living. Table #2.3.12 gives the cost of living for the U.S. from the years 1947 through 2011, with the year 1977 being used as the year that all others are compared (DeNavas-Walt, Proctor & Smith, 2012). Create a time-series plot and interpret.

Table #2.3.12: Data of Time versus CPI

Year	CPI-U-RS1 index (December 1977=100)	Year	CPI-U-RS1 index (December 1977=100)
1947	37.5	1980	127.1
1948	40.5	1981	139.2
1949	40.0	1982	147.6
1950	40.5	1983	153.9
1951	43.7	1984	160.2
1952	44.5	1985	165.7
1953	44.8	1986	168.7
1954	45.2	1987	174.4
1955	45.0	1988	180.8
1956	45.7	1989	188.6
1957	47.2	1990	198.0
1958	48.5	1991	205.1
1959	48.9	1992	210.3
1960	49.7	1993	215.5
1961	50.2	1994	220.1
1962	50.7	1995	225.4
1963	51.4	1996	231.4
1964	52.1	1997	236.4
1965	52.9	1998	239.7
1966	54.4	1999	244.7
1967	56.1	2000	252.9
1968	58.3	2001	260.0
1969	60.9	2002	264.2
1970	63.9	2003	270.1
1971	66.7	2004	277.4
1972	68.7	2005	286.7
1973	73.0	2006	296.1
1974	80.3	2007	304.5
1975	86.9	2008	316.2
1976	91.9	2009	315.0
1977	97.7	2010	320.2
1978	104.4	2011	330.3
1979	114.4		

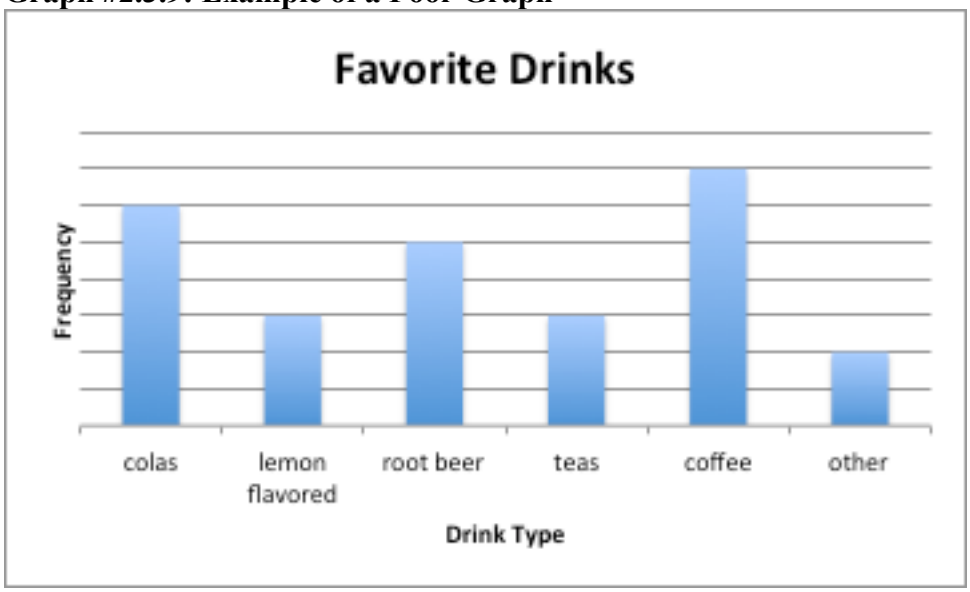
10.) The median incomes for all households in the U.S. for the years 1967 to 2011 are given in table #2.3.13 (DeNavas-Walt, Proctor & Smith, 2012). Create a time-series plot and interpret.

Table #2.3.13: Data of Time versus Median Income

Year	Median Income	Year	Median Income
1967	42,056	1990	49,950
1968	43,868	1991	48,516
1969	45,499	1992	48,117
1970	45,146	1993	47,884
1971	44,707	1994	48,418
1972	46,622	1995	49,935
1973	47,563	1996	50,661
1974	46,057	1997	51,704
1975	44,851	1998	53,582
1976	45,595	1999	54,932
1977	45,884	2000	54,841
1978	47,659	2001	53,646
1979	47,527	2002	53,019
1980	46,024	2003	52,973
1981	45,260	2004	52,788
1982	45,139	2005	53,371
1983	44,823	2006	53,768
1984	46,215	2007	54,489
1985	47,079	2008	52,546
1986	48,746	2009	52,195
1987	49,358	2010	50,831
1988	49,737	2011	50,054
1989	50,624		

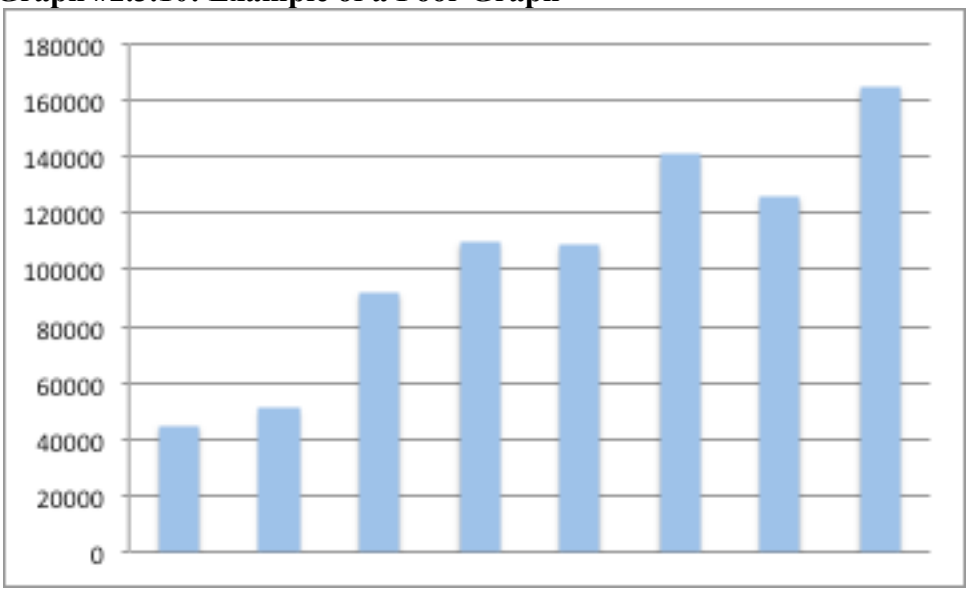
11.) State everything that makes graph #2.3.9 a misleading or poor graph.

Graph #2.3.9: Example of a Poor Graph



12.) State everything that makes graph #2.3.10 a misleading or poor graph (Benen, 2011).

Graph #2.3.10: Example of a Poor Graph

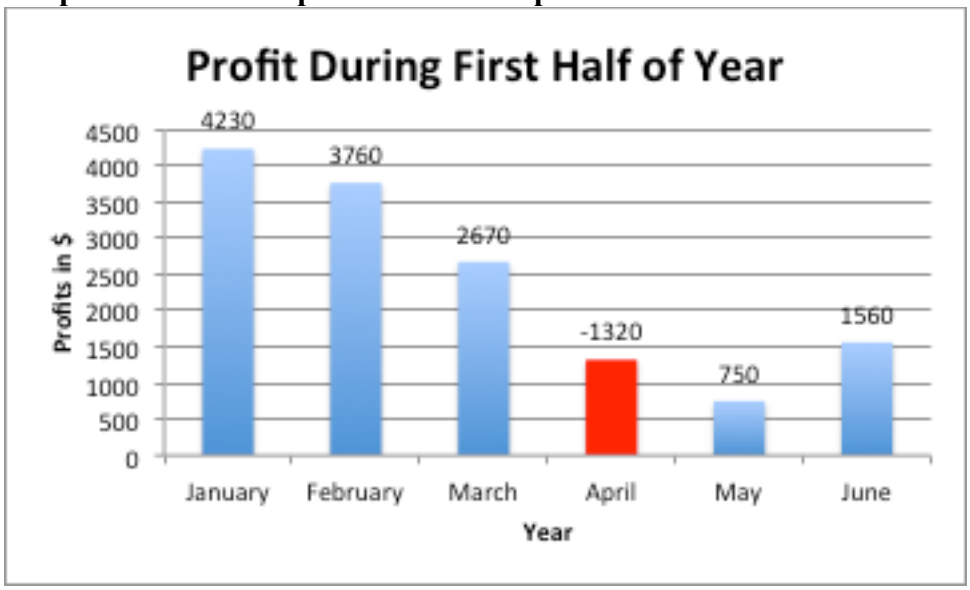


13.) State everything that makes graph #2.3.11 a misleading or poor graph ("United States unemployment," 2013).

Graph #2.3.11: Example of a Poor Graph

14.) State everything that makes graph #2.3.12 a misleading or poor graph.

Graph #2.3.12: Example of a Poor Graph



Data Sources:

B1 assets of financial institutions. (2013, June 27). Retrieved from www.rba.gov.au/statistics/tables/xls/b01hist.xls

Benen, S. (2011, September 02). [Web log message]. Retrieved from http://www.washingtonmonthly.com/political-animal/2011_09/gop_leaders_stop_taking_credit031960.php

Capital and rental values of Auckland properties. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/rentcap.html

Contraceptive use. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=35

Deaths from firearms. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/firearms.html

DeNavas-Walt, C., Proctor, B., & Smith, J. U.S. Department of Commerce, U.S. Census Bureau. (2012). *Income, poverty, and health insurance coverage in the United States: 2011* (P60-243). Retrieved from website: www.census.gov/prod/2012pubs/p60-243.pdf

Density of people in Africa. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=30&loc=249,250,251,252,253,254,34227,255,257,258,259,260,261,262,263,264,265,266,267,268,269,270,271,272,274,275,276,277,278,279,280,281,282,283,284,285,286,287,288,289,290,291,292,294,295,296,297,298,299,300,301,302,304,305,306,307,308

Department of Health and Human Services, ASPE. (2013). *Health insurance marketplace premiums for 2014*. Retrieved from website: http://aspe.hhs.gov/health/reports/2013/marketplacepremiums/ib_premiumslandscape.pdf

Electricity usage. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=162

Fertility rate. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.TFRT.IN

Fuel oil usage. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=164

Gas usage. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=165

Health expenditure. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS

Hinatov, M. U.S. Consumer Product Safety Commission, Directorate of Epidemiology. (2012). *Incidents, deaths, and in-depth investigations associated with non-fire carbon monoxide from engine-driven generators and other engine-driven tools, 1999-2011*. Retrieved from website: http://www.cpsc.gov/PageFiles/129857/cogenerators.pdf

Life expectancy at birth. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

Median income of males. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=137

Median income of males. (2013, October 9). Retrieved from http://www.prb.org/DataFinder/Topic/Rankings.aspx?ind=136

Prediction of height from metacarpal bone length. (2013, September 26). Retrieved from http://www.statsci.org/data/general/stature.html

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.STA.ANVC.ZS

United States unemployment. (2013, October 14). Retrieved from http://www.tradingeconomics.com/united-states/unemployment-rate

Weissmann, J. (2013, March 20). A truly devastating graph on state higher education spending. *The Atlantic*. Retrieved from http://www.theatlantic.com/business/archive/2013/03/a-truly-devastating-graph-on-state-higher-education-spending/274199/