DATA DISPLAY

Documents prepared for use in course B01.1305,
New York University, Stern School of Business

Cover: Clydesdale horses, Martha Clara Vineyards,
Riverhead, New York, July 2005

Revision date 8 JAN 2007

≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈DATA DISPLAY≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈

⋘⋘⋘⋘⋘⋘⋘⋘⋘⋘⋘⋘DATA TYPES⋙⋙⋙⋙⋙⋙⋙⋙⋙⋙⋙⋙

A number of distinctions are made with regard to data. These are easy to understand through examples.

Univariate or *multivariate*?

Univariate information consists of one fact for each object in the data base. The listing of the rates of returns for each of 500 mutual funds is univariate information.

Multivariate information consists of two or more facts for each object in the data base. The listing of (rates of return, load fees, management overhead) for each of 500 mutual funds is multivariate information.

Multivariate information may of course be reduced to univariate information. For example, you could deal with rates of return (only) even though you also have other information.

Univariate information is sometimes called *scalar*, and multivariate information is sometimes called *vector*.

Qualitative or *quantitative*?

Quantitative information is presented as numbers permitting arithmetic. Qualitative information is everything else. This sounds easy, but watch out for non-numeric information tagged with numbers. For example, consider a set of responses about laundry detergents:

1 = Tide
2 = Surf
3 = Wisk
4 = Cheer
(etc)

This information is qualitative, and the numbers are mere labels of convenience. In particular, you cannot do arithmetic. The average of 1 and 3 is 2, but you cannot say that the average of Tide and Wisk is Surf.

Information is quantitative whenever arithmetic (such as taking averages) makes sense.

For quantitative data only, *discrete* or *continuous*?

This is an easy distinction if you think of
discrete = counted
continuous = measured

Data obtained by counting are almost always integers, and there are no values between. In noting the numbers of children of the employees of a business, there are no values between 2 and 3. Measured data allow these “in-between” values. Between 62 inches and 63 inches are many possible values.

> There is a corresponding grammatical distinction between the words *fewer* and *less*. The word *fewer* is used with things that are counted, as in “There were fewer orders placed last month.” The word *less* is used with things that are measured, as in “I had less sugar than I needed, so I couldn’t make the cookies.”

Sometimes continuous, or measured, data appears to be discrete because it is crudely rounded. For example, human heights rounded to the nearest centimeter should be considered as measured even though the values are integers like 140, 141, 142, 143, …. Ages should also be considered as measured even though they are reported as integers. It is an obfuscation to say that ages are obtained by “counting birthdays” — age data should still be treated as continuous.

What is *ordinal* data?

While quantitative data always possesses an ordering, qualitative data may or may not have an ordering. The questionnaire categories disagree strongly, disagree, neutral, agree, agree strongly form ordinal information. These are sometimes called Likert scales. We can list these in the order

disagree strongly < disagree < neutral < agree < agree strongly

The reverse order would also make sense, but any scrambled order would look silly. Suppose that we endow these categories with a set of numbers:

1 = disagree strongly
2 = disagree
3 = neutral
4 = agree
5 = agree strongly

Even in this form, the information would be regarded as qualitative. In some situations, analysts will treat these as quantitative (and find, say, the average of a set of responses). If you do want to think of these as quantitative, please be aware that the scores are subjective. The numbers 1, 2, 3, 4, 5 are operationally equivalent to the numbers -20, -10, 0, 10, 20 because the relative spacings are maintained. The numbers 1, 2, 5, 6, 8 have different relative spacings, and would not be considered equivalent to 1, 2, 3, 4, 5.

For quantitative data, we often make a further distinction between *interval* data and *ratio* data. This distinction is rarely important, but we'll mention it here anyhow.

Interval information allows for the calculation of meaningful differences. For example, Celsius temperatures are interval data; the difference between 10º and 20º is the same as the difference between 50º and 60º.

Ratio information allows divisions. For example, an object weighing 60 kg is 20% heavier than an object weighing 50 kg. That is, we can calculate
(60-50) ÷ 50 = 20%.

Celsius temperatures do not have the ratio property. We cannot say that 60º is 20% hotter than 50º.

Ratio information implies the existence of a meaningful *zero* point. We cannot say that 60º is 20% hotter than 50º. The temperature zero point is artificial.

Most variables — heights, weights, values, times, pressures, and so on — have the zero property. Indeed, ratio data is the most common form.

⋘⋘⋘⋘⋘⋘⋘⋘⋘⋘DATA TYPES⋙⋙⋙⋙⋙⋙⋙⋙⋙⋙

The major reason for recognizing the types of data is that certain types of arithmetic are to be avoided for some of them.

> Qualitative data allows you to count responses.
>
> Ordinal data allows you to count responses and find medians, quartiles and other percentiles. (If these data are endowed with a numerical scale, admittedly artificial, then you can compute means, medians, standard deviations, and so on.)
>
> Quantitative (interval or ratio) data allows you to find medians, quartiles and other percentiles, means, and standard deviations.
>
> Quantitative ratio data allows all the calculations used on quantitative interval data, along with things like the coefficient of variation (which depends on a zero point).

In multivariate information, the data can be of different types. For example, in a listing of facts about municipal bonds, we could encounter (municipality name, interest rate, bond rating). These three are nominal, ratio, and ordinal.

Time series or *cross sectional* ?

> Data in which time sequencing is a relevant factor are regarded as time series. A string of daily stock prices, a list of annual economic data over 20 years, or a human subject’s blood pressure collected daily for a month should all be regarded as time series. When such data are presented in spreadsheet format, the actual dates form one of the columns.
>
> Data that are not time series are *cross-sectional*. All the data values are contemporaneous, obtained at the same time.
>
> It is possible for a set of data to have both time series and cross-sectional aspects. As an example, consider a data base involving 200 subjects and containing each subject’s height, weight, marital status, and daily blood pressure values for one month.

⋘⋘⋘⋘⋘⋘⋘⋘PARETO CHART ⋙⋙⋙⋙⋙⋙⋙⋙

Consider a set of qualitative values. A simple display device is the Pareto chart, a naive display device requiring that you sort the values by frequency of occurrence. For instance, a frozen pizza baker might note sales by types:

plain, mushroom, pepper and onion, plain, pepperoni, pepperoni, plain, …

The simple summary will consist of the counts of each type over some specified sales window (such as one week). It might look like this:

Pizza type	Number sold
Pepperoni	2,428
Plain	3,615
Mushroom	1,800
Sausage	645
Pepper and onion	810
Mushroom and onion	1,025
Garlic	253
Meatball and sausage	558

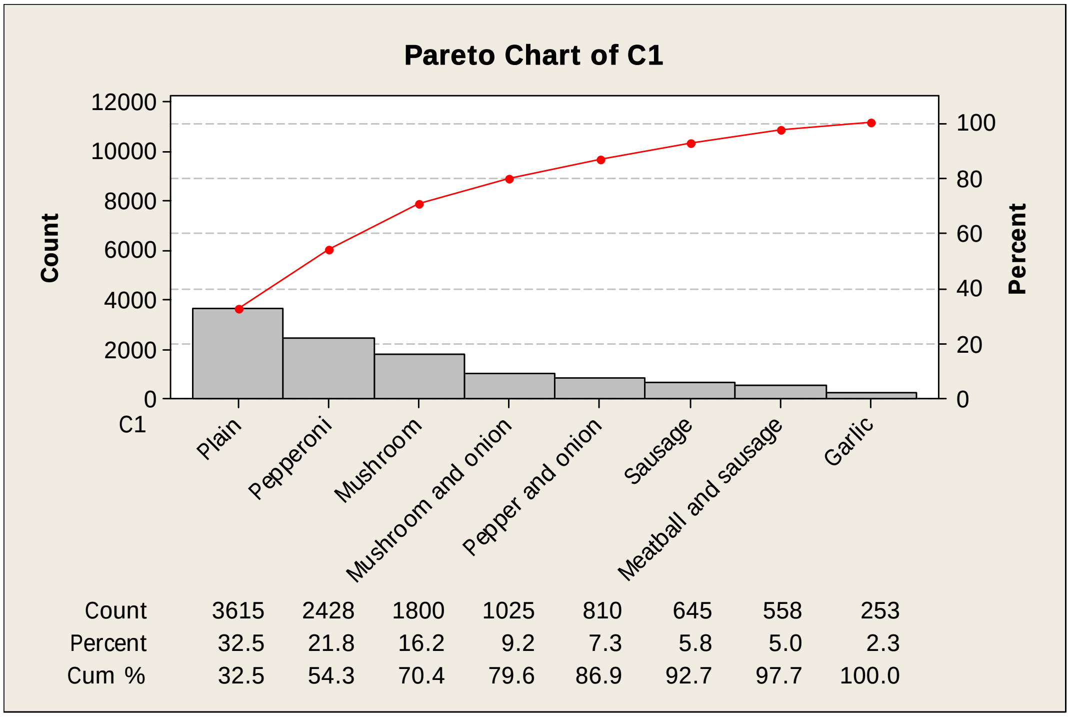
One should note immediately that the total number of pizzas sold was 11,134. This will allow you to express the pizza types in percentages, as follows:

Pizza type	Number sold	Percent
Pepperoni	2,428	21.8
Plain	3,615	32.5
Mushroom	1,800	16.2
Sausage	645	5.8
Pepper and onion	810	7.3
Mushroom and onion	1,025	9.2
Garlic	253	2.3
Meatball and sausage	558	5.0
TOTAL	11,134	100.1

The Pareto chart sorts these by popularity: plain, pepperoni, mushroom, … These are then displayed as bars along with a curve giving the cumulative percentages. This task is easy enough by hand, but computer assistance will allow you to make a much more attractive picture.

The following was obtained in Minitab, through **Stat** ⇒ **Quality Tools** ⇒ **Pareto Chart** ⇒.

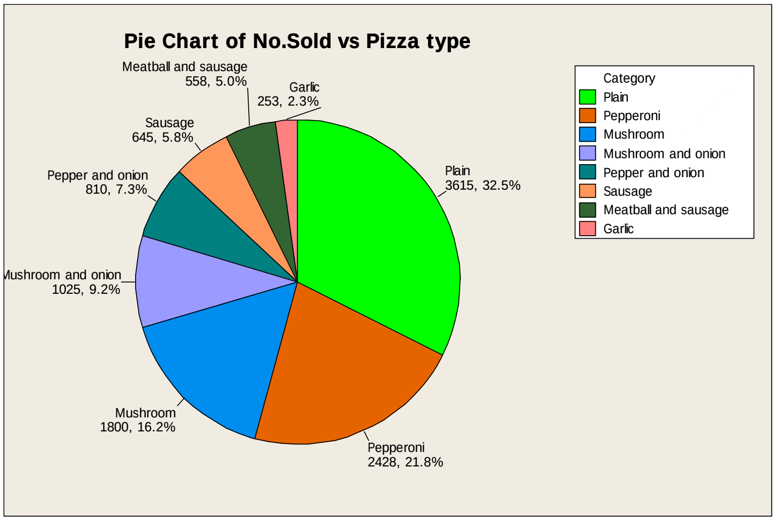
⋘⋘⋘⋘⋘⋘⋘⋘PARETO CHART ⋙⋙⋙⋙⋙⋙⋙⋙



Count	3615	2428	1800	1025	810	645	558	253
Percent	32.5	21.8	16.2	9.2	7.3	5.8	5.0	2.3
Cum %	32.5	54.3	70.4	79.6	86.9	92.7	97.7	100.0

This is not a very sophisticated statistical notion. It is, however, very easy to understand.

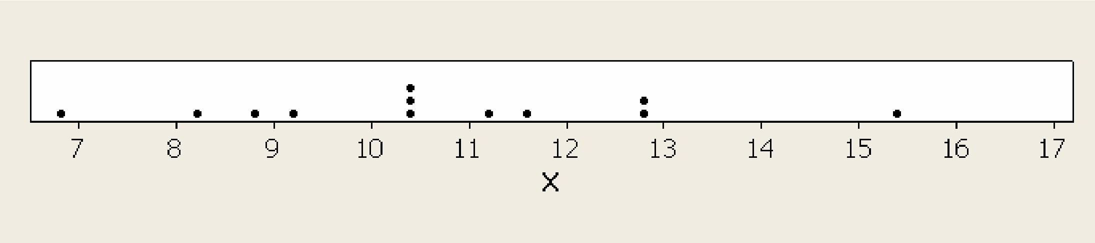
A common alternative to a Pareto chart is a pie chart. Minitab will do this with **Graph ⇒ Pie Chart**. Selecting the options to label the slices will give this:



In general we cannot recommend pie charts. It is visually difficult to compare angles, and it is especially hard to make comparisons on side-by-side pie charts.

⋘⋘⋘⋘⋘⋘⋘⋘DOT DIAGRAM⋙⋙⋙⋙⋙⋙⋙⋙

The dot diagram is sometimes used as a visual display to show a set of data. Consider, for instance, the display below:

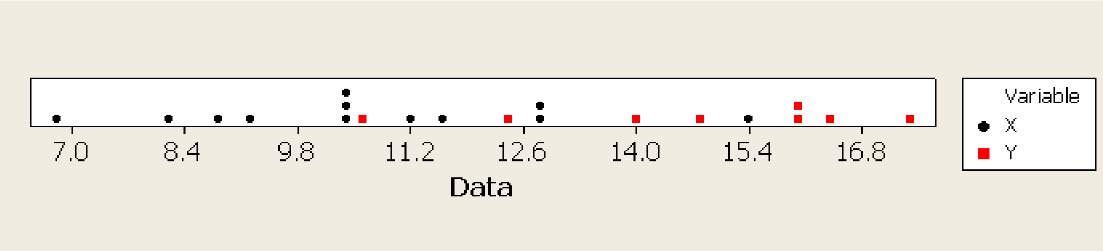


This shows a set of data with values (in order)

6.8 8.2 8.8 9.2 10.4 10.4 10.4 11.2 11.6 12.8 12.8 15.4

The usefulness of the dot diagram is limited to small sets of data. It helps that the data values in this example are crudely rounded. You'd have some difficult decisions to make if the values reported here as 10.4 were actually 10.364, 10.392, 10.438. This plot was made with the Minitab sequence **Graph** ⇒ **Dotplot**.

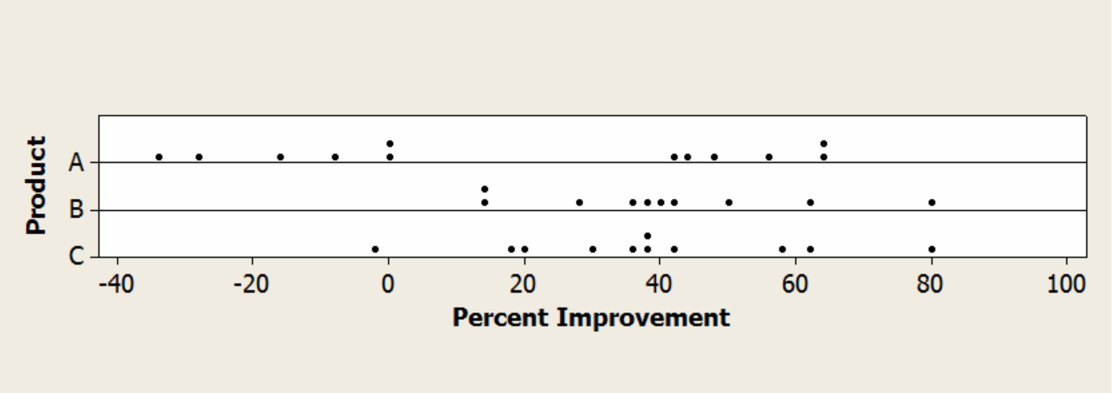
One can sometimes use this diagram to compare two or more sets of data. This requires that the two sets be identified by different symbols.



This was made by Minitab 14, using **Graph** ⇒ **Dotplot**, and selecting **Multiple Y's, Stack Y's**.

In this example, it is clear that the group represented by the squares takes values which tend to be larger than those for the group with circles.

Displays like the previous are improved when the groups are separated on different axes. With Minitab 14, use **Graph** ⇒ **Dotplot**, and then select **One Y**, **With Groups**. This display shows the responses of subjects to three different anti-arthritis medications.



There are some clear limitations. These diagrams get very cluttered with even medium sample sizes.

⋘⋘⋘⋘⋘⋘THE STEM-AND-LEAF-DISPLAY⋙⋙⋙⋙⋙⋙

Let’s consider what you’ve got to do to summarize a set of data. The 250 values below represent the percent of account value in equities at a certain brokerage firm. For example, the first value 83.5 represents the fact that in the first sampled account, 83.5% o the value was in equities. The remaining 16.5% was in some combination of bonds, treasury securities, precious metals, cash, real estate, or other instruments.

We’d like to summarize this list of values.

83.5	67.4	85.6	94.9	72.3	52.9	59.3	71.4	62.4	72.0
71.2	63.8	59.2	73.7	86.7	83.6	64.2	59.2	48.4	59.7
57.4	88.2	77.2	43.0	84.3	89.1	65.8	75.6	73.8	69.0
61.0	63.2	67.7	59.1	64.4	88.2	79.5	81.4	59.5	38.9
94.5	73.7	92.0	93.2	87.6	31.8	73.0	67.4	70.2	69.5
39.7	71.0	85.9	99.0	63.8	79.5	96.9	71.9	61.6	42.1
77.7	54.6	63.6	85.6	67.3	83.8	70.0	67.7	73.1	51.5
76.4	75.6	65.2	71.7	83.8	75.5	46.5	63.9	80.0	39.9
65.5	82.4	80.1	85.0	77.8	81.2	64.1	52.5	58.2	65.4
90.6	65.3	65.1	70.2	84.0	54.8	62.7	87.6	74.8	72.7
40.3	63.8	45.6	54.8	65.9	64.3	87.8	63.1	55.1	51.8
45.8	75.8	79.5	52.7	42.3	71.8	56.3	66.6	94.9	53.3
72.9	65.2	67.3	57.8	73.2	78.4	74.6	61.2	82.2	84.9
57.7	36.6	76.7	95.2	61.2	69.2	92.5	65.0	61.4	46.4
94.6	95.7	47.2	60.4	34.3	66.2	83.6	58.6	61.4	64.3
38.5	86.7	66.8	78.5	56.1	73.4	83.0	37.7	90.7	70.2
43.1	54.2	79.1	95.5	93.7	51.2	72.5	69.9	35.1	74.9
66.8	73.7	63.9	55.0	36.7	55.0	60.1	46.6	69.1	67.5
39.6	68.4	62.9	65.0	54.4	65.5	78.2	62.5	66.2	79.5
63.0	87.2	48.2	56.8	48.1	74.1	36.8	56.8	77.3	31.0
54.1	78.7	94.7	72.9	57.5	78.5	87.9	57.9	56.2	72.1
61.4	48.8	67.4	67.3	68.8	68.9	55.7	63.8	50.8	90.8
69.2	80.6	68.5	78.4	45.1	57.8	68.1	67.4	72.8	52.9
69.2	78.8	75.6	66.2	96.2	70.6	86.0	55.8	87.6	37.5
90.3	74.2	70.0	53.9	39.9	84.0	79.3	43.7	58.4	54.7

One device for summarizing such data is the *frequency distribution*. This device is made by setting up categories in a framework such as the following:

Category	Frequency
Less than 35.0%	
35.0% to 54.9%	
55.0% to 74.9%	
75.0% to 94.9%	
95.0% and higher	

If you wish to make a frequency distribution, you will have to decide the number of categories to use, the widths of the categories, and the division points. You’ll also have to go through the annoying clerical steps of counting up the numbers in each of the categories. Generally this is done by going through the list and placing a tick mark for each entry. Thus, after going through the first five values, the work would look like this:

Category	Frequency
Less than 35.0%	
35.0% to 54.9%	
55.0% to 74.9%	| |
75.0% to 94.9%	| | |
95.0% and higher	

At the completion of the counting, the tick marks are replaced by numbers.

Though you will see frequency distributions now and then, the whole idea is somewhat obsolete, and we have better techniques. Minitab, for example, does not provide frequency distributions. Minitab does provide stem-and-leaf displays, and you can convert a stem-and-leaf display into a frequency distribution.

Let's begin the discussion of the stem-and-leaf by showing that it can be constructed by hand. Let's note that the leading digit, here the *tens* digit, can be an integer from 0 to 9. We'll set up a display listing these from small to large:

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

Now let's note that the first value on our list, 83.5, has a tens digit of 8 and a units digit of 3. We'll decide to forget about the third digit, avoiding even the issue of rounding off. We locate *tens* = 8, *units* = 3 as follows:

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

We next locate 67.4 (*tens* = 6, *units* = 7) and 85.6 (*tens* = 8, *units* = 5) to produce this:

```
0 |
1 |
2 |
3 |
4 |
5 |
6 | 7
7 |
8 | 35
9 |
```

After the first ten values (the first row), we have this:

```
0 |
1 |
2 |
3 |
4 |
5 | 29
6 | 72
7 | 212
8 | 35
9 | 4
```

After the first thirty values (the first three rows), the display has grown to be this:

```
0 |
1 |
2 |
3 |
4 | 43
5 | 299997
6 | 723459
7 | 21213753
8 | 3563849
9 | 4
```

In this device, the value to the left of the vertical bar is called the stem, and the values to the right are called the leaves.

If the data set ended at n = 30, we'd sort the leaves, trim off unused stem positions at the high and low ends, and add a legend. This would give

```
4 | 34
5 | 279999
6 | 234579          NOTE:  6 | 3  means 63
7 | 11223357
8 | 3345689
9 | 4
```

The legend is important, as we need to indicate that the stem values are multiples of 10. In other data sets, the stem values might be 10,000 or 100 or 0.0001.

The device is called a stem-and-leaf display. It was invented by John Tukey of Princeton University and Bell Labs. You might think about the following issues.

In making the stem-and-leaf display, there are subjective choices to be made. For example, if the list of values had gone from 418.2 to 1,272.5 the best stem choice would have been 100s, and the stem values would have been 4 5 6 7 8 9 10 11 12.

We disregard digits below the precision of the leaf digits. Proper rounding would be a distracting nuisance. Moreover, we down-bias the operation only by one-half the value of the leaf value. In the example above, this bias would be $\frac{1}{2} \times 1 = 0.5$. Disregarding the low-order digits has another interesting advantage; in our example, any value x in the row beginning `6 |` has the property $60 \le x < 70$. If we had done proper rounding the condition would be the rather awkward property $59.5 \le x < 69.5$.

The stem-and-leaf display can be made in a one-pass operation. That is, you have to handle each value exactly once in doing this.

Amazingly little information is lost in going from the original data values to the stem-and-leaf display.

The process of creating the stem-and-leaf display automatically sorts the values. Specifically, after you've done the stem-and-leaf display, the median value is very easy to locate.

Some lists of values do not make nice stem-and-leaf displays. For example, the list 25.4, 1388, 0.5, 16, 105, 27218, … runs through several orders of magnitudes. As a plausible option, you might replace the values by their logarithms before doing the stem-and-leaf display (or other statistical work).

Some people like to put gaps at certain positions to make the leaves easier to count. It's hardly necessary for our example, but here is the display with gaps after every fifth leaf:

```
4 | 34
5 | 27999 9
6 | 23457 9          NOTE:  6 | 3  means 63
7 | 11223 357
8 | 33456 89
9 | 4
```

The stem-and-leaf display is easily converted to a histogram. Just place bars over the picture!

```
40├──┐
  │34│
50├──┴───┐
  │279999│
60├──────┤
  │234579│
70├──────┴─┐
  │11223357│
80├───────┬┘
  │3345689│
90├─┬─────┘
  │4│
  ├─┘
  │
```

Clearly you’ve got to avoid gaps between the leaves. You’ve also got to print the leaves in some equi-size font. You want skinny 1s to take up as much space as fat 8s. In this picture, the Courier font was used. You’ve also got to deal with some decisions about how this should be labeled. Our choice here put values at the hash marks between bars.

Finally, let’s note that this project is tedious if the sample size is large. For our full data set of n = 250, we can use Minitab’s **Graph** ⇒ **Stem-and-Leaf** ⇒ option. (This is also available as **Stat** ⇒ **EDA** ⇒ **Stem-and-Leaf** ⇒.) This will give us the following:

```
Stem-and-leaf of C1        N  = 250
Leaf Unit = 1.0

    3    3 114
   15    3 566677889999
   21    4 022333
   32    4 55566678888
   49    5 01112222334444444
   74    5 5555566666777777888999999
  102    6 0011111112222333333333344444
 (39)    6 555555555556666666777777777888889999999
  109    7 0000001111112222222223333333344444
   76    7 5555566777788888888999999
   52    8 00011223333334444
   35    8 5555666777777889
   19    9 0000223344444
    6    9 555669
```

Here C1 identifies the Minitab column in which the values were placed. Minitab reminds us that the sample size is n = 250. We are also told that the leaf unit is 1; this implies that the stem unit is 10, so that the list begins 31, 31, 34, 35, 36, 36, ..

Minitab has made the decision to split the stem positions. That is, a single row for *stem* = 3 has been replaced by two rows. These two rows are (*stem* = 3, *leaf* = 0,1,2,3,4) and (*stem* = 3, *leaf* = 5,6,7,8,9). Sometimes the stem positions are split into five parts.

The values in the left column are counts from the outside. The last row of the display has 6 values. The last two rows together have 19 values, the last three rows together have 35 values, and so on. The counting is done toward the row which contains the median. Minitab notes that row by showing in parentheses the count in that row. The row marked (39) has 39 values and it contains the median.

Since there are 250 values, the median occurs between positions 125 and 126. The rows preceding the median row have a cumulative count of 102. The 125th value must be at position 125 - 102 = 23 within this row. This value is one of the 67s; the 126th value is also one of the 67s, so we report the median as 67.

> If you are obsessed with pointless precision, you would worry that the low-order digits have been ignored in construction of the stem-and-leaf display. You could go back to the original data list and go through the process of sorting out the values between 67 and 68 to get a more precise median.

⋘⋘⋘⋘⋘⋘⋘⋘THE BOXPLOT⋙⋙⋙⋙⋙⋙⋙⋙

The simple unadorned boxplot shows just five facts: the minimum, the lower quartile (lower hinge), the median, the upper quartile (upper hinge), and the maximum. This simple style has some obvious advantages:

It's easy to create by hand.
It's easy to explain.

There are, however, some drawbacks to the simple boxplot:

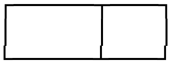
It fails to satisfy curiousity about very large and very small data values.

The long whiskers visually over-emphasize a trivial aspect of the data.

Accordingly, there is a more detailed version of the boxplot. It gives plenty of information, but it is somewhat difficult to construct without computer assistance. Curiously, there does not really exist an extra explaining burden for the detailed boxplot; these things are easy to read even if one does not understand all the nuances of their construction.

Here then is the building plan for the detailed boxplot. Remember, you rarely have to build any... you just need to be able to read them.

The central part of the boxplot is unchanged. That is, you still show a central box involving the lower quartile, the median, and the upper quartile. The picture at this stage (with a number line beneath) might look like this:



```
--+----+----+----+----+----+----+----+----+----+----+----+----+----+--
 100  200  300  400  500  600  700  800  900 1000 1100 1200 1300 1400 $
```

The number line suggests that we are dealing with dollar-value information.

The picture shows (at least within drawing precision) that LQ = Q_1 = lower quartile = 25th percentile = \$400, median = \$520, and UQ = Q_3 = upper quartile = 75th percentile = \$600.

Now determine IQR = interquartile range = UQ - LQ = Q_3 - Q_1 = \$600 - \$400 = \$200.

Next find the following quantities:

UIF = upper inner fence = UQ + 1.5 × IQR

= \$600 + 1.5 × \$200 = \$900

UOF = upper outer fence = UQ + 3 × IQR

= \$600 + 3 × \$200 = \$1,200

LIF = lower inner fence = LQ - 1.5 × IQR

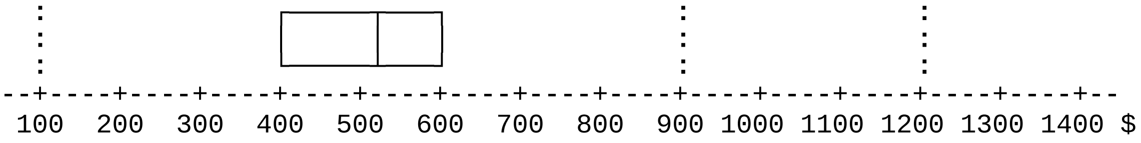
= \$400 - 1.5 × \$200 = \$100

LOF = lower outer fence = LQ - 3 × IQR

= \$400 - 3 × \$200 = -\$200

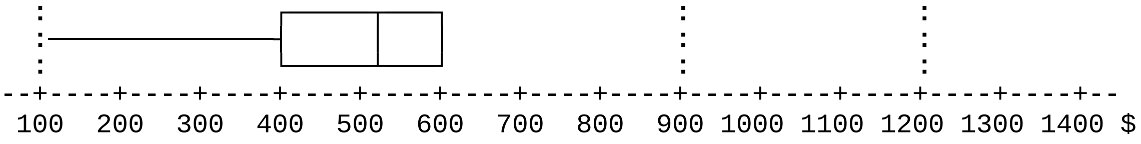
The values “1.5” and “3” have become traditional in these definitions.

Here now is the boxplot with the fences. This picture is *not* part of the final display.



(The lower outer fence is not shown, as it is off the scale. If this lower outer fence were relevant to this set of values, we’d have to revise the scale.)

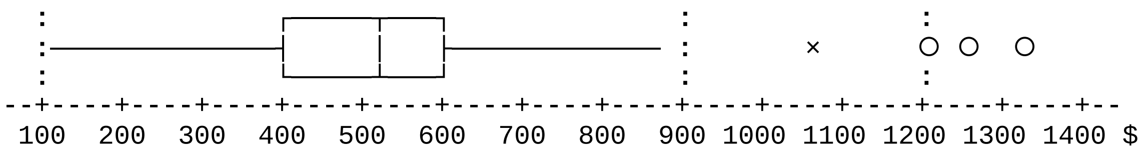
Examine now the minimum value in the list. Suppose that this value is \$120. This value is between the inner lower fence and the lower quartile and is thus nonremarkable. Thus we make an ordinary whisker at the lower end of the box, stretching down to the \$120 value. Our picture has now evolved to this:



Now look at the maximum value in the list. Suppose that this value is $1,320. This is above the upper inner fence and worthy of special note. Indeed, we now need to look down from the large values until we get to a value inside the fences. Suppose that the largest values in the list (in decreasing order) are these:

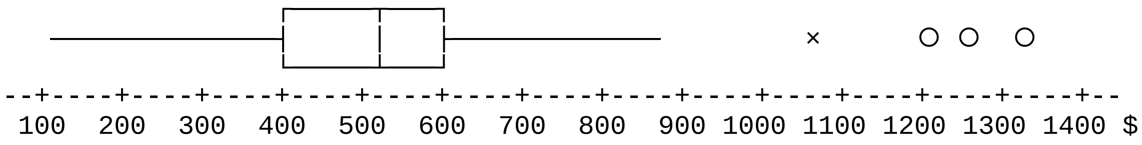
1,320 1,240 1,200 1,060 860 …

The values outside the fences are sometimes called *outliers* and get special marks. We then complete the picture by drawing a whisker from the upper quartile to 860, which is the most remote non-outlier. The next picture is this:



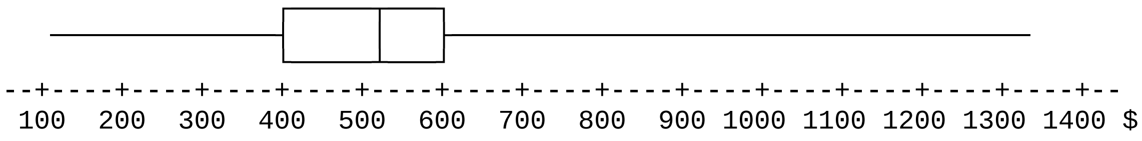
We used one symbol, here ×, for values between the inner and outer fences, and we used another, here ○, for values beyond the outer fence.

The final activity is to erase the fences. This gives us our final picture:



In fact, it wasn't necessary to draw in the fences in the first place!

For the sake of comparison, the simple boxplot for these numbers would be this



Observe that the detailed boxplot is easy to read, in spite of the complexities involved in its construction.

Most of the time our boxplots are computer-generated, so that we don’t have to play with the construction details. In dealing with your clients, you don’t want to have to explain the “fences” concepts; it is easy enough to say that the × and ○ symbols represent unusual individual values.

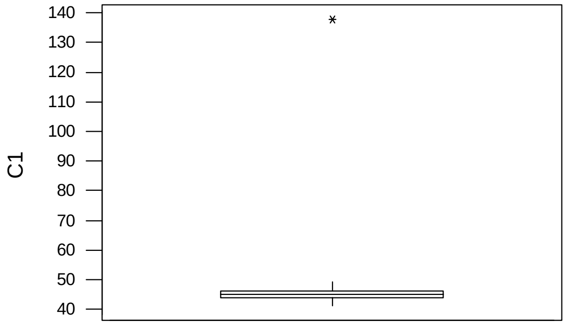
Some fine points:

> Values outside of fences are implicitly defined as *outliers* in this method. This is not a universal definition.
>
> Many embellishments have been proposed for the boxplot, but these have not proved popular. Such variations include tapering for the boxes, notches at the centers of the boxes, using the box width to express the sample size, and so on.
>
> The boxplot is one of the display devices of EDA, or exploratory data analysis. EDA is the creation of John Tukey of Princeton University and Bell Labs.

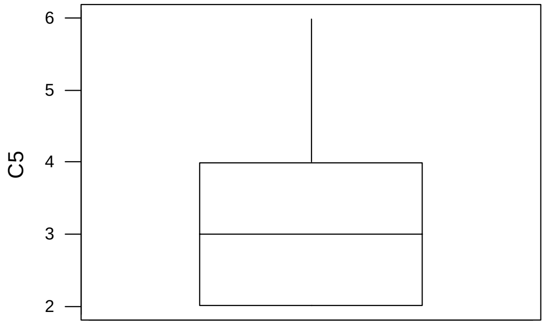
⋘⋘⋘⋘⋘⋘⋘⋘BOX PLOT PATHOLOGIES ⋙⋙⋙⋙⋙⋙⋙⋙

Box plots can show unusual pathologies. This document shows several of these. The stories are given after the fourth picture.

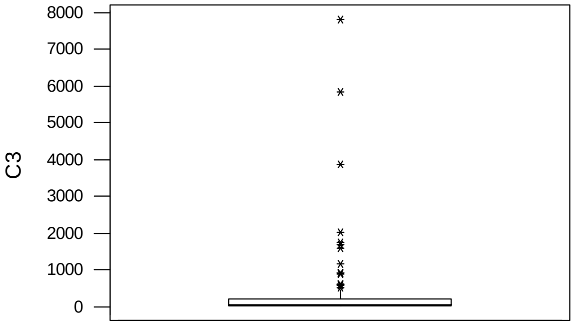
(1) Consider this:



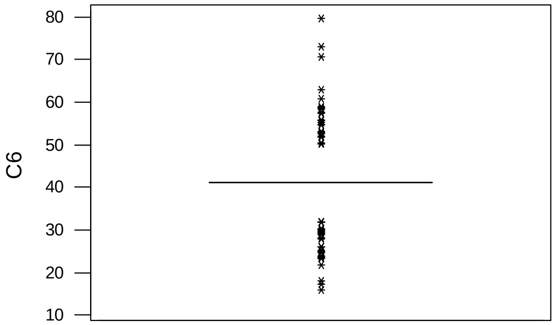
(2) The box plot below lacks a lower whisker. How could this happen?



(3) There is a serious problem here:



(4) What kind of data could have created this?



⋘⋘⋘⋘⋘⋘⋘⋘BOX PLOT PATHOLOGIES ⋙⋙⋙⋙⋙⋙⋙⋙

(1) This is the result of one extremely unusual data value.   This data value would qualify as an outlier by almost any definition. It should be checked.

If correct, it should be set aside and discussed separately.

If incorrect but repairable, it should be corrected.

If incorrect and not repairable, it should be removed.

(2) This picture tells us that Minimum = Q_L . Apparently there are a lot of tied values.

(3) The long string of high-side outliers suggests that these data are extremely positively skewed. (The data will behave much more reasonably if they are replaced by their logarithms.) You might notice that the box has very little detail; this is due to extensive ties among the values.

(4) The box part of this box plot has been reduced to nothing but a line! This can only happen with a very extensive tied set of values in the middle of the data. It happens that IQR = interquartile range = 0, so that every point not equal to the median is an outlier!