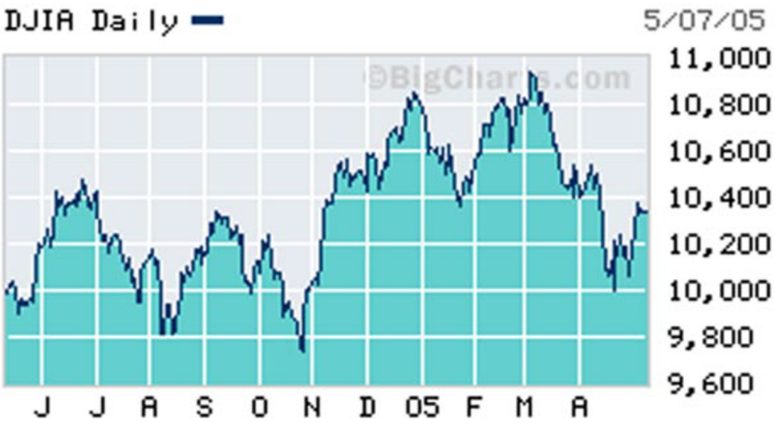
2. Graphs



One of the most powerful ways to summarize data is a graph. A graph is a visual tool that lets the reader quickly discern trends and make comparisons.

The graph above is from MS MONEY and summarizes a one-year history of the Dow Jones Industrial Average. What do you think of this graph?

The `GRAPHS.XLSX` spreadsheet accompanies this section.

2.1. Histograms

Suppose that you administer an IQ test to 128 high school students and obtain the scores at right.
Since the scores are sorted from lowest to highest, we can see that they range from 57 to 141. What other patterns can you see from this list of scores?
There are so many scores there is not very much **information.**

57	84	90	96	100	104	106	114
65	84	90	96	100	104	107	117
67	84	90	97	101	104	107	117
71	84	92	97	101	104	107	118
76	84	92	97	101	104	109	119
76	85	93	98	101	104	109	120
76	85	93	98	102	105	110	121
77	85	93	98	102	105	111	124
78	86	93	99	102	105	111	125
78	86	94	99	102	106	111	127
79	86	94	99	102	106	112	127
80	87	95	99	103	106	112	129
82	88	95	99	103	106	112	131
83	88	95	99	103	106	112	133
83	88	95	100	103	106	113	133
83	89	95	100	103	106	113	141

Our goal is to summarize the data in order to deduce patterns. Right now we cant see the forest for the leaves.

We will construct ranges or cells and count how many actual observations fall in each cell.
The purpose of this is to help us visualize what information is in the data.
The range of our observations is

$$141 - 57 = 84$$

If we wanted, say, exactly ten cells, then we would need cell to be 8.4 units long. But our goal is not to be exact but instead to visualize the data. So, instead of using 8.4 for each cell, round 8.4 to a more convenient number, say 10. Will we get exactly ten cells using cell lengths of ten? Does it matter?
Next decide the bottom – lower limit – for the first cell. For example, we could make the lower limit for the first cell 40.

A better choice might be 56 (so that the upper limit of the cell is 65, ending in 5).
This results in cells that look like

lower	upper	count
56	65	
66	75	
76	85	
86	95	
96	105	
106	115	
116	125	
126	135	
136	145	

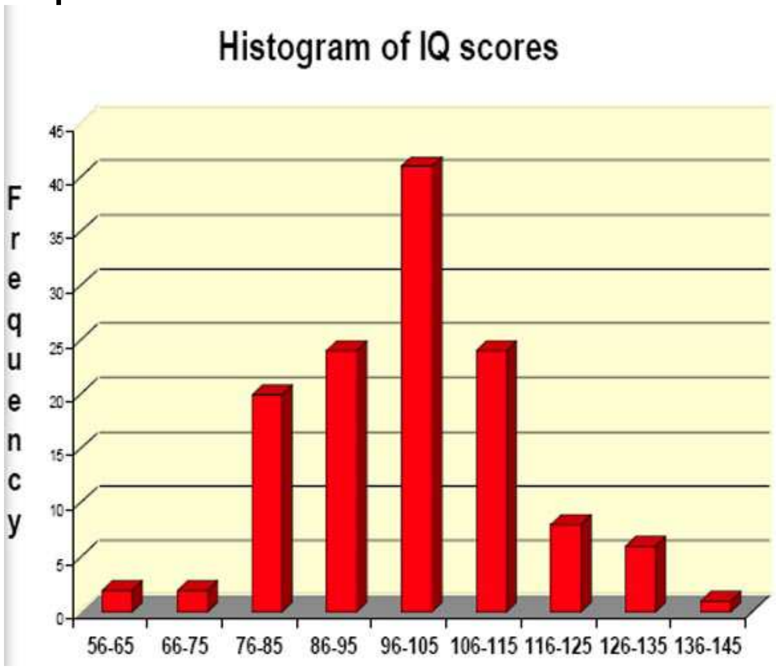
There are of course other possible choices–this is just one of many possibilities.

Filling out the table then gives:

lower	upper	count
56	65	2
66	75	2
76	85	20
86	95	24
96	105	41
106	115	24
116	125	8
126	135	6
136	145	1

57	84	90	96	100	104	106	114
65	84	90	96	100	104	107	117
67	84	90	97	101	104	107	117
71	84	92	97	101	104	107	118
76	84	92	97	101	104	109	119
76	85	93	98	101	104	109	120
76	85	93	98	102	105	110	121
77	85	93	98	102	105	111	124
78	86	93	99	102	105	111	125
78	86	94	99	102	106	111	127
79	86	94	99	102	106	112	127
80	87	95	99	103	106	112	129
82	88	95	99	103	106	112	131
83	88	95	99	103	106	112	133
83	88	95	100	103	106	113	133
83	89	95	100	103	106	113	141

The final step is to construct a bar chart or histogram that graphically represents the data



The height of each bar corresponds to the count in each category.

Some other observations that are now apparent:

- The observations appear to have a “bell-shaped” distribution.
- Most of the observations are in the middle range. In fact, 89.5% of the observations fall between 86 and 115:

$$89.5\% = \frac{24 + 41 + 24}{128} \times 100\%$$

- The observations appear to be symmetrically distributed, centered roughly at 100.

Fortunately, Excel makes it easy to do histograms, provided that the Data Analysis Tool tool has been installed. While this is a standard component of Excel, it does not install by default. Once installed, it's easy to use.

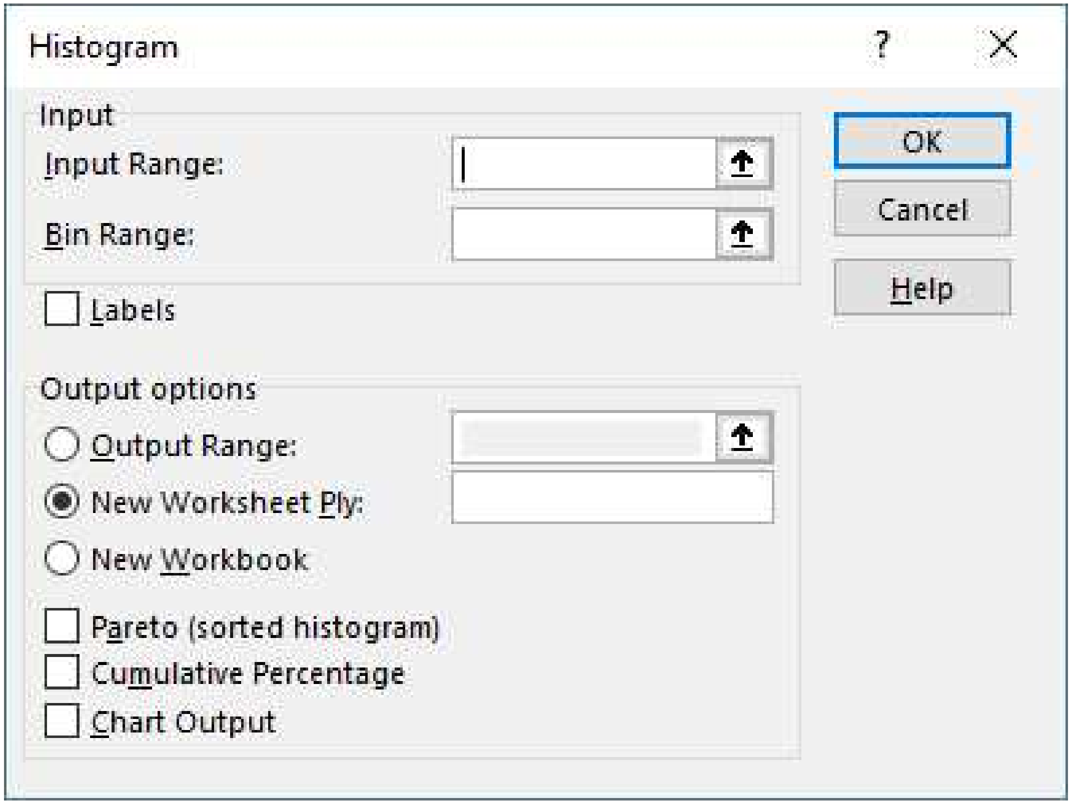
Solution Template

Step 1. Enter your data in a spreadsheet. The data do not have to be ordered.

Step 2. Figure out your bins–these are the intervals into which you will sort your data.

Step 3. Each bin has a lower and upper limit. Enter the upper limits into the spreadsheet.

Step 4. Use `Data –> Data Analysis –> Histogram` to create your histogram.



Histogram ? ×

Input

Input Range:

Bin Range:

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

□ Chart Output

In the tool, fill in the cell ranges for your data and your bins. You can optionally choose a location for the output. Check `Chart Output` to have it produce a chart.

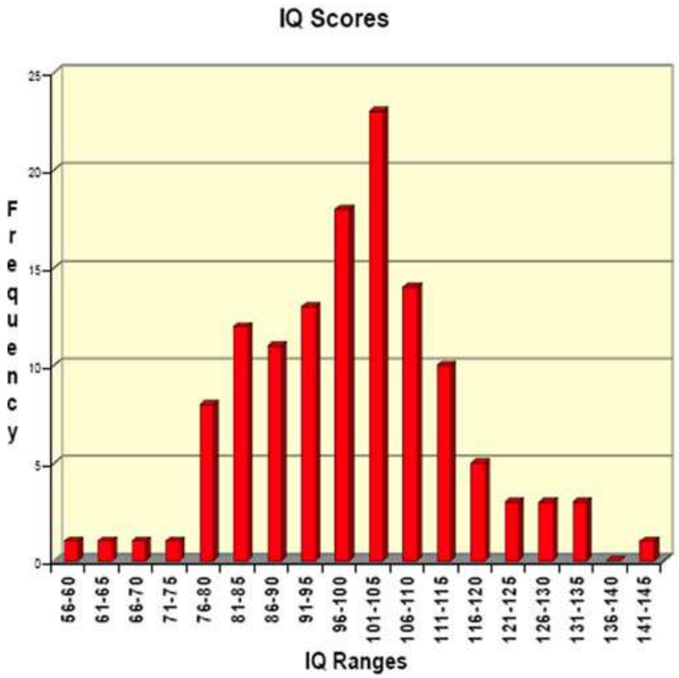
Step 5. Construct the histogram.

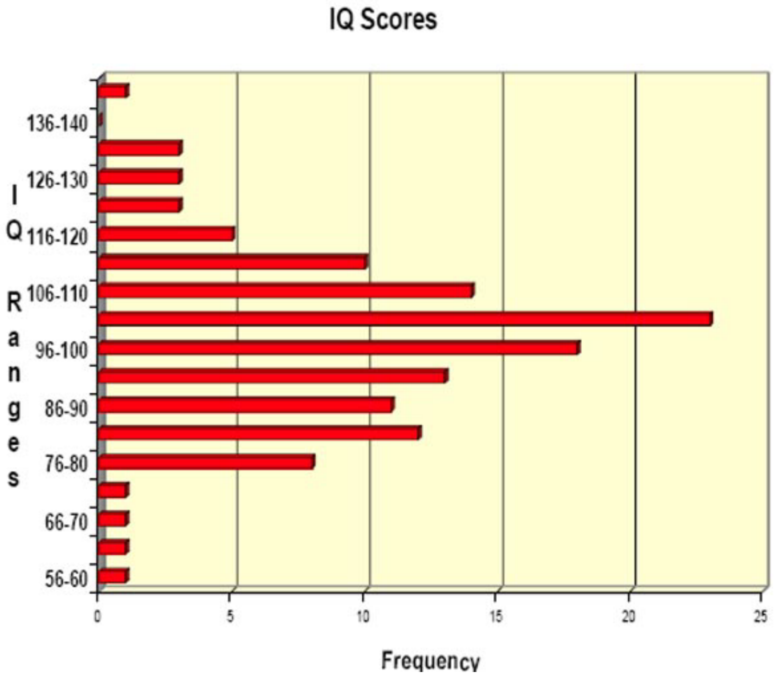
End of Solution Template

You do it.

Starting with the same data as in the previous example, construct a histogram with approximately fifteen cells.

57	84	90	96	100	104	106	114
65	84	90	96	100	104	107	117
67	84	90	97	101	104	107	117
71	84	92	97	101	104	107	118
76	84	92	97	101	104	109	119
76	85	93	98	101	104	109	120
76	85	93	98	102	105	110	121
77	85	93	98	102	105	111	124
78	86	93	99	102	105	111	125
78	86	94	99	102	106	111	127
79	86	94	99	102	106	112	127
80	87	95	99	103	106	112	129
82	88	95	99	103	106	112	131
83	88	95	99	103	106	112	133
83	88	95	100	103	106	113	133
83	89	95	100	103	106	113	141





Which histogram is easier to read?

2.2. Nominative Scales

There are four kinds of scales, or ways of measuring subjects, used in behavorial sciences:

- Nominative scale
- Ordinal scale
- Interval scale
- Ratio scale

Nominative or naming scales assign labels to subjects, usually based on attributes. These can divide subjects into categories, such as gender, place of birth, or some other non-quantitive characteristic. Sometimes nominative scales involve numbers, but they are used as labels. For example, offensive and defensive linemen on a football team wear uniforms numbered in the 60s or 70s, while quarterbacks wear uniforms numbered between 1 and 19. The relative value of that number has no meaning other than as a label.

An ordinal scale is similar to a nominative scale, in that it divides the subjects into categories. However, ordinal scales include the notion of ordering. Examples might be grades, socioeconomic status, or rank.

An interval scale is similar to nominative and ordinal scales, but the magnitudes between adjacent intervals are the same. Temperature measured in degrees Fahrenheit or Celsius is an example. Clearly ninety degrees is hotter than fifty degrees (ordering), and the difference between forty degrees and fifty degrees is the same as the difference between twenty degrees and thirty degrees. Interval scores don't have a true zero, although there may be an artificial zero such as the freezing point of water. Time and date are interval scales. Interval scales may or may not be infinitely divisible, so GRE scores are interval scores.

Finally, ratio scales add a true zero to the mix. Physical measurements like height, weight, blood pressure, and distance are ratio scales. Elapsed time which measures the time from a particular event has a true zero. The zero in a ratio scale represents the absence of what is being measured. Thus, temperature in degrees Kelvin which uses absolute zero–the total absence of heat–is a ratio scale. GRE scores are not ratio scales since all GRE scores range between 200 and 800–there is no zero.

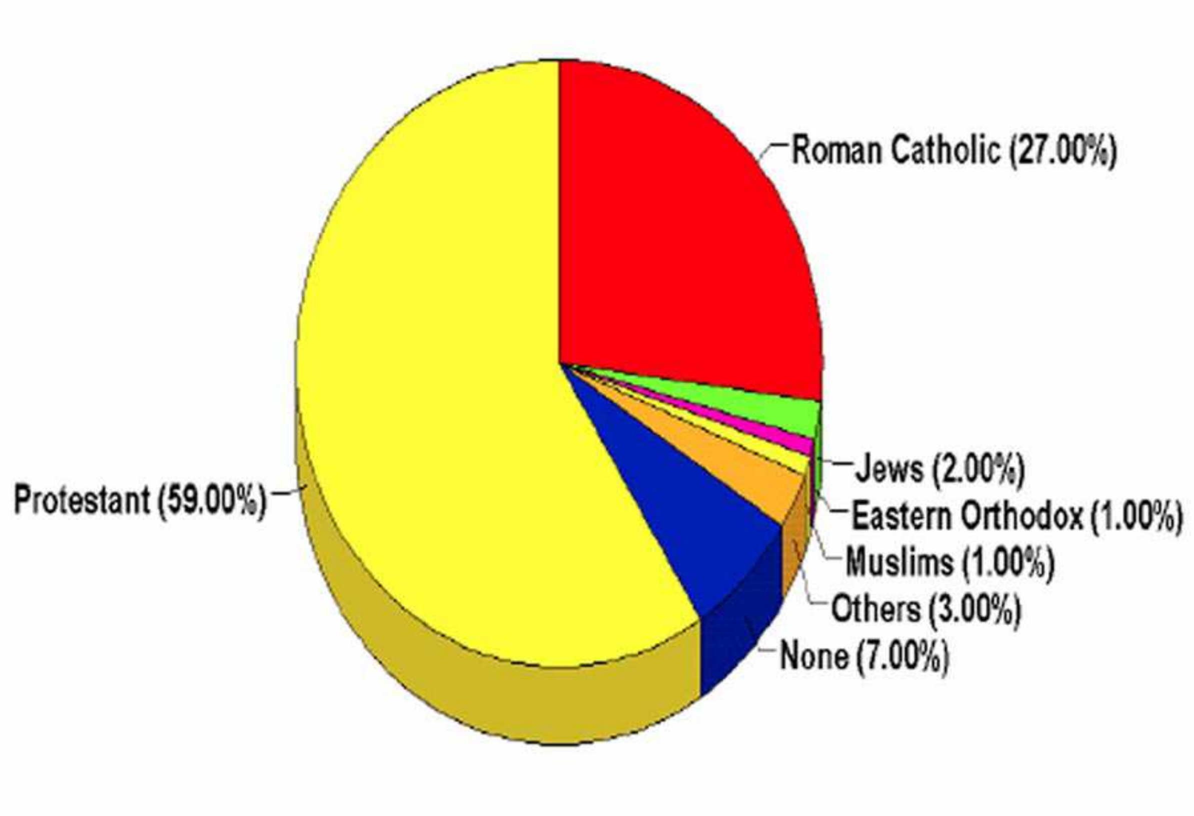
Because interval and ratio scales involve magnitudes, they are quantitative scales, while nominative and ordinal scales are qualitative or attribute scales. Quantitative and attribute scales are summarized in different ways.

Pie charts are the appropriate visual presentation for nominative scales and for some ordinal scales.

For example, a recent poll asked the respondents to self-identify their religious affiliation. The results were as follows:

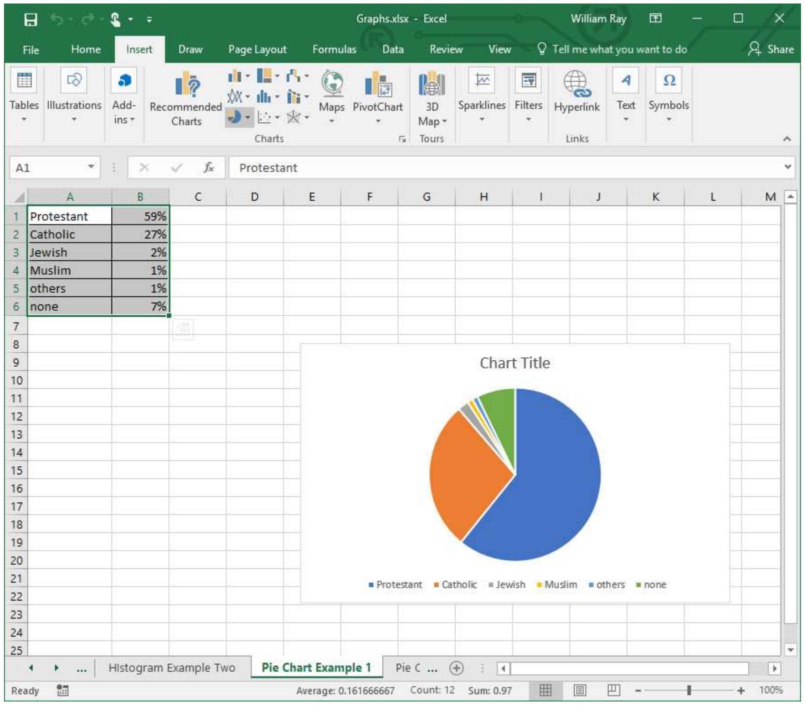
Protestant	59%
Catholic	27%
Jewish	2%
Muslim	1%
others	1%
none	7%

The resulting pie chart is:



Who invented pie charts?

Excel makes it easy to produce pie charts, too. First, create a table that includes your categories. Then highlight the table and `INSERT` –> `CHARTS` and chose a pie chart.



	A	B
1	Protestant	59%
2	Catholic	27%
3	Jewish	2%
4	Muslim	1%
5	others	1%
6	none	7%

A1 | Protestant

Ready | Average: 0.161666667 | Count: 12 | Sum: 0.97 | 100%

For another example, research suggests that there are three factors that contribute to long-term well-being or happiness:

Environmental factors	10%
Genetic factors	50%
Intential activities	40%

We can easily use Excel to make a pie chart of the above table.

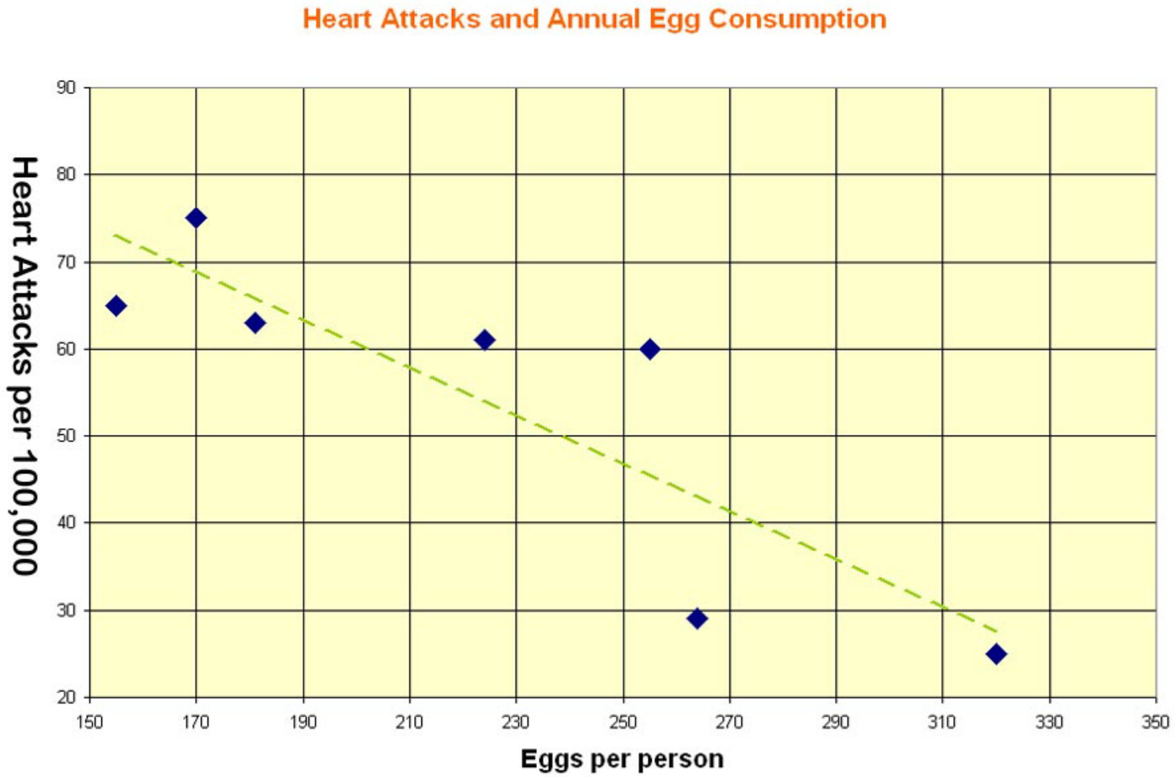
2.3. Scatter Plots

Sometimes you will gather two or more quantitative measures on each subject. In these cases you are often interested in determining if there is a relationship between the two variables (height and weight for example). The visual presentation that can help you understand this is the scatter plot.

Suppose for example you gather data on annual egg consumption and annual heart attack rates for several countries:

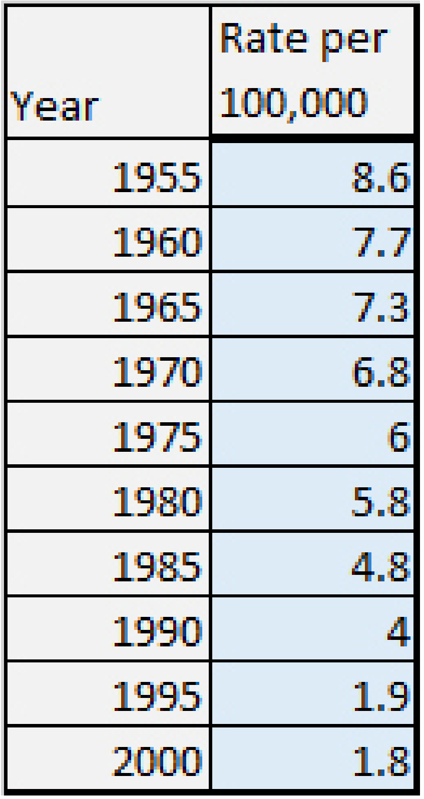
Country	Annual Eggs	Mortality
Australia	155	65
UK	170	75
Canada	181	63
Germany	224	61
US	255	60
France	264	29
Japan	320	25

A scatter plot for this data is



Note that we added a trend line to this scatter plot. What conclusions might you draw from this plot?

Excel makes it easy to do scatter plots, too. For another example, consider the following data on deaths in the workplace per 100,000 workers.



Year	Rate per 100,000
1955	8.6
1960	7.7
1965	7.3
1970	6.8
1975	6
1980	5.8
1985	4.8
1990	4
1995	1.9
2000	1.8