4 Charts, Histograms, Errors in Graphing

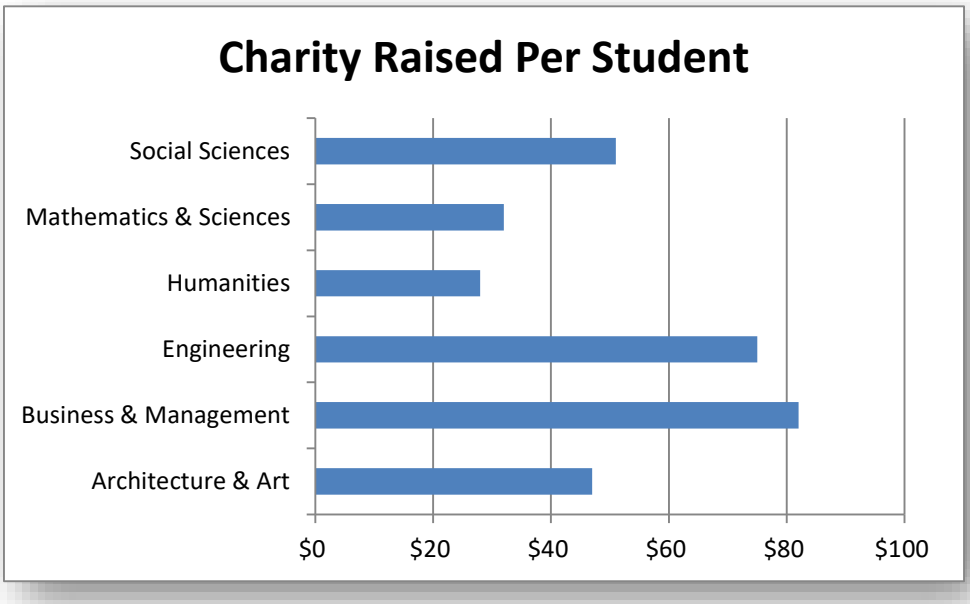
4.1 Horizontal Bar Charts vs Vertical Column Charts

When we need to think about and represent data sets involving categorical and ordinal data, we should use horizontal bar charts rather than vertical bar charts. Vertical bar charts are, by convention, usually reserved for interval and ratio scales. For example, consider the made-up data in about the average amount of charity donations raised by students in different schools of a university in Figure 4-1. A horizontal bar chart (Figure 4-2) represents the charity raised per student by the length of the horizontal bar corresponding to each school.

Figure 4-1. Donations per student by school.

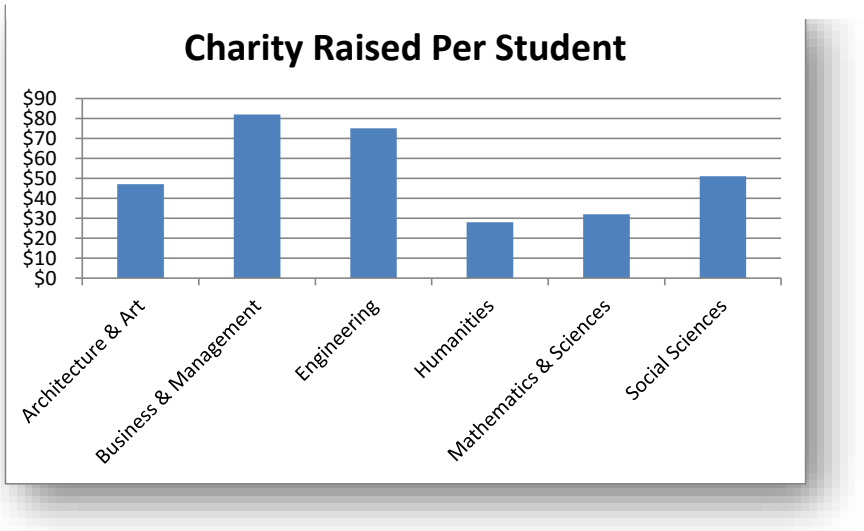
School	Charity Raised Per Student
Architecture & Art	$47
Business & Management	$82
Engineering	$75
Humanities	$28
Mathematics & Sciences	$32
Social Sciences	$51

Figure 4-2. Horizontal bar chart showing donations by school.



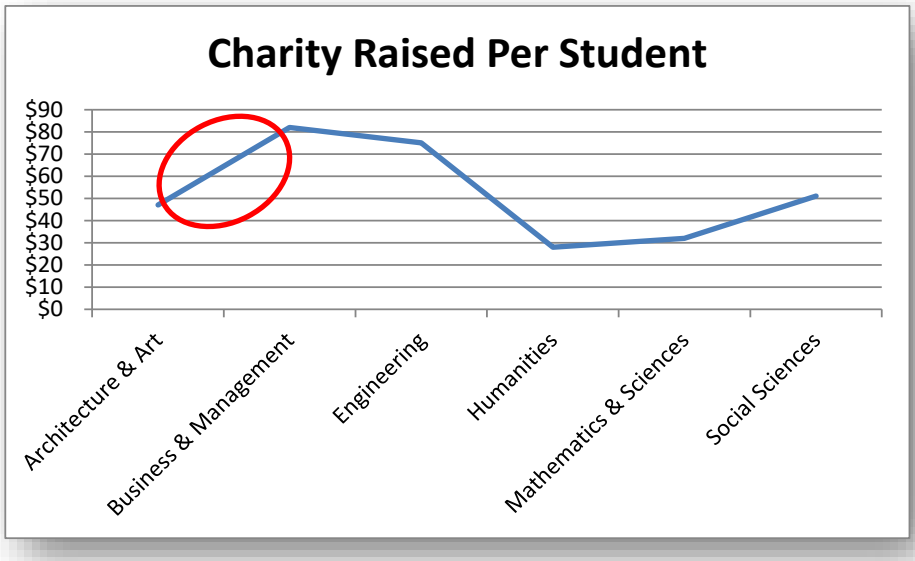
The advantage of horizontal bar charts versus vertical charts is that there is *no implication* that the axis with the category labels (in this case, the vertical axis) represents any kind of quantitative variable. In contrast, putting the bars vertically is (by a widely flouted convention) an indication that there is *some sort of quantitative relationship* among the labels of the bars. So although the same information *can* be represented as a vertical bar chart (Figure 4-3), some experts argue that the hidden assumption behind any charts that show the main variable(s) on the vertical axis (the *ordinate* or *Y-axis*) is that the horizontal axis (the *abscissa* or *X-axis*) is a continuous variable, providing some meaning to values between the observed markers.

Figure 4-3. Vertical bar (column) chart showing donations by school.



To illustrate how misleading it would be to put categorical variables on the X-axis, Figure 4-4 shows the same data as Figure 4-2 and Figure 4-3 but using a *line graph*. Everyone will agree that the values between the observed points (see red oval) make no sense at all: there are no values of the X-axis (donations per student) between values of the X-axis (school; e.g., *Architecture & Art* and *Business and Management*). Figure 4-4 is a definite no-no.

Figure 4-4. Bad choice of representation for categories.



4.2 Pie Charts

Another way of representing information about categories is the pie chart. Generally, a pie chart shows segments whose areas add up to a meaningful total and where the area of a segment (or what's equivalent, the angle of the vertex of the pie) corresponds proportionally to the value for that segment. Figure 4-5 shows the total charitable donations per school in a single semester at a university.

Figure 4-7 shows the pie graph for these data arranged with the segments in alphabetical order. Each segment has a different color and is identified in the legend (the list of schools with a small colored square before each school).

Figure 4-5. Total donations per school in one semester.

School	Total Donations
Architecture & Art	$4,606
Business & Management	$33,046
Engineering	$23,401
Humanities	$9,792
Mathematics & Sciences	$6,560
Social Sciences	$31,263

One of the conventions that *may* be used in a pie chart is that the data are sorted by size and the largest component segment is placed starting at the 12:00 position, with the decreasing segments placed in clockwise order. Figure 4-6 shows the same data arranged in this conventional pattern.

Figure 4-7. Pie chart showing total donations per school.

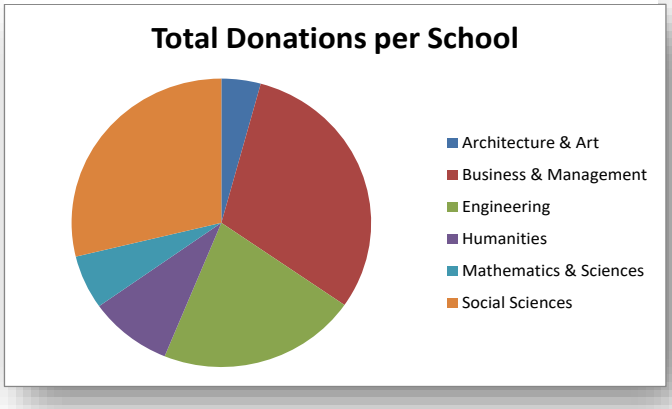
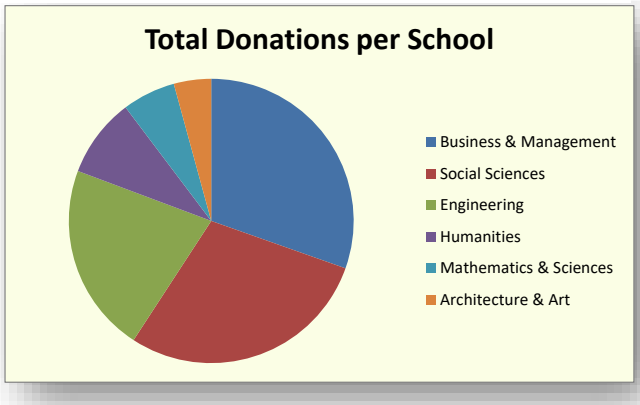


Figure 4-6. Pie-chart using clockwise pattern of descending values.



To change chart type, right-click on the pie and use the Change Chart Type menu shown in Figure 4-8:

Figure 4-8. EXCEL menu for changing chart type.

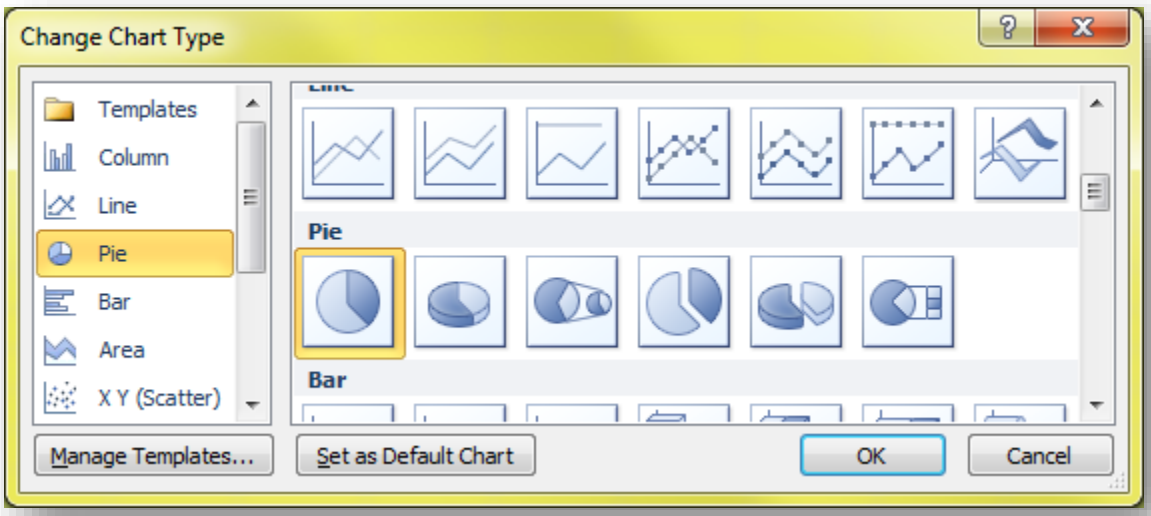
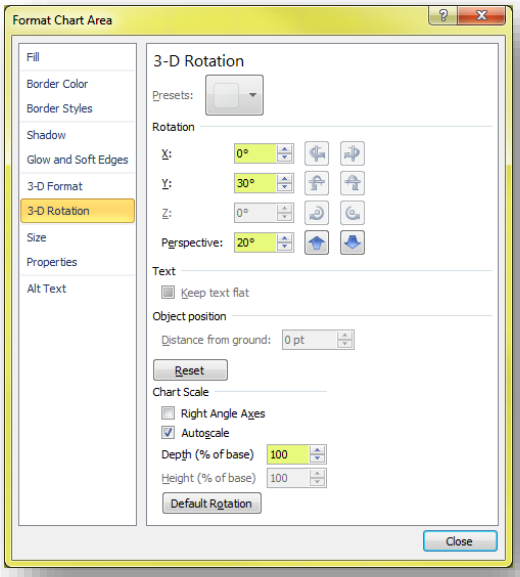
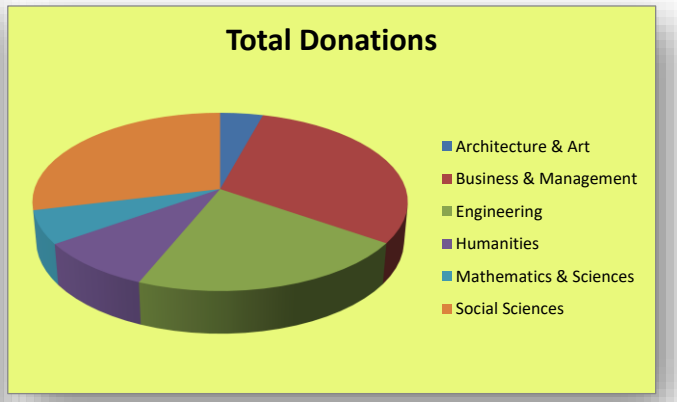


Figure 4-9 shows the result of converting Figure 4-7 to a 3D pie chart and then applying the 3-D Rotation options.

Figure 4-9. Menu and results for changing from pie chart to 3D pie chart and applying 3D rotation.





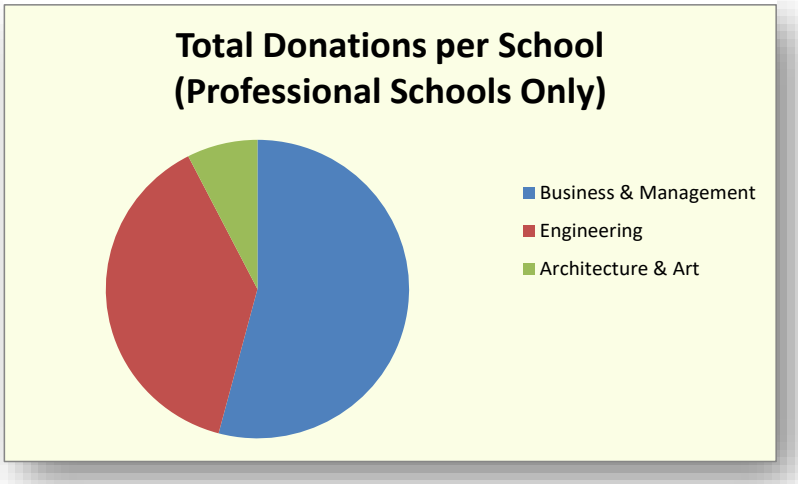
Be careful in using the Perspective angle – a high angle can distort the appearance of the segments, increasing the apparent size of the closer segments.

INSTANT TEST P 4

Create a pie chart with some interesting data. Change the style to 3D and experiment with various rotation angles in the X and Y axes using the 3-D Rotation menu. Try out options in the 3-D Format menu. Play around with other options such as Border Color, Border Styles, Shadow, and Glow and Soft Edges.

The total for a pie chart does not necessarily have to be the sum of *all* possible values in a table; it may be reasonable in specific cases to create a chart showing the relative proportions of *selected* components. For example, someone might be interested primarily in discussing the donations of students in the professional schools in the data of Figure 4-5; a pie chart for those selected data would look like Figure 4-10.

Figure 4-10. Total Donations per School (selected schools only).

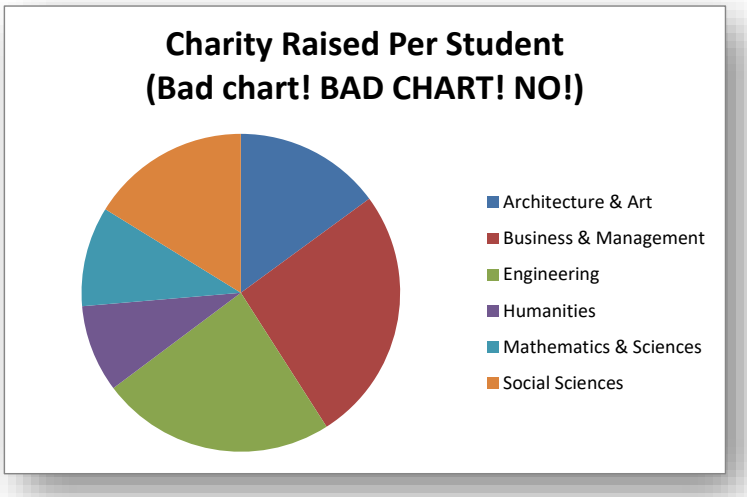


IMPORTANT WARNING ABOUT BAD PIE CHARTS

Don't create pie charts for data you cannot legitimately add up.

For example, Figure 4-11 is a nonsensical chart that shows donations *per student* for each of the schools. Those numbers are *averages*. **You can't add up averages unless the number of data points (usually the *sample size*) are the same for all the averages.** If the School of Architecture & Art has half the number of students that the School of Business & Management has, it doesn't make sense to add up their averages. So a pie chart of averages would not make sense, since one does not add averages up together – one computes the total from multiplying the average by the number of observations (students, here) that gave rise to the average. Figure 4-11 is an example of a *bad pie chart* that doesn't make sense.

Figure 4-11. Pie chart created using averages (BAD).



INSTANT TEST P 5

Explain as if to a smart youngster exactly *why* it does not make sense to create a pie chart based on averages.

Use examples of real or invented data to illustrate your explanation.

4.3 Clustered and Stacked Bar Charts and Column Charts

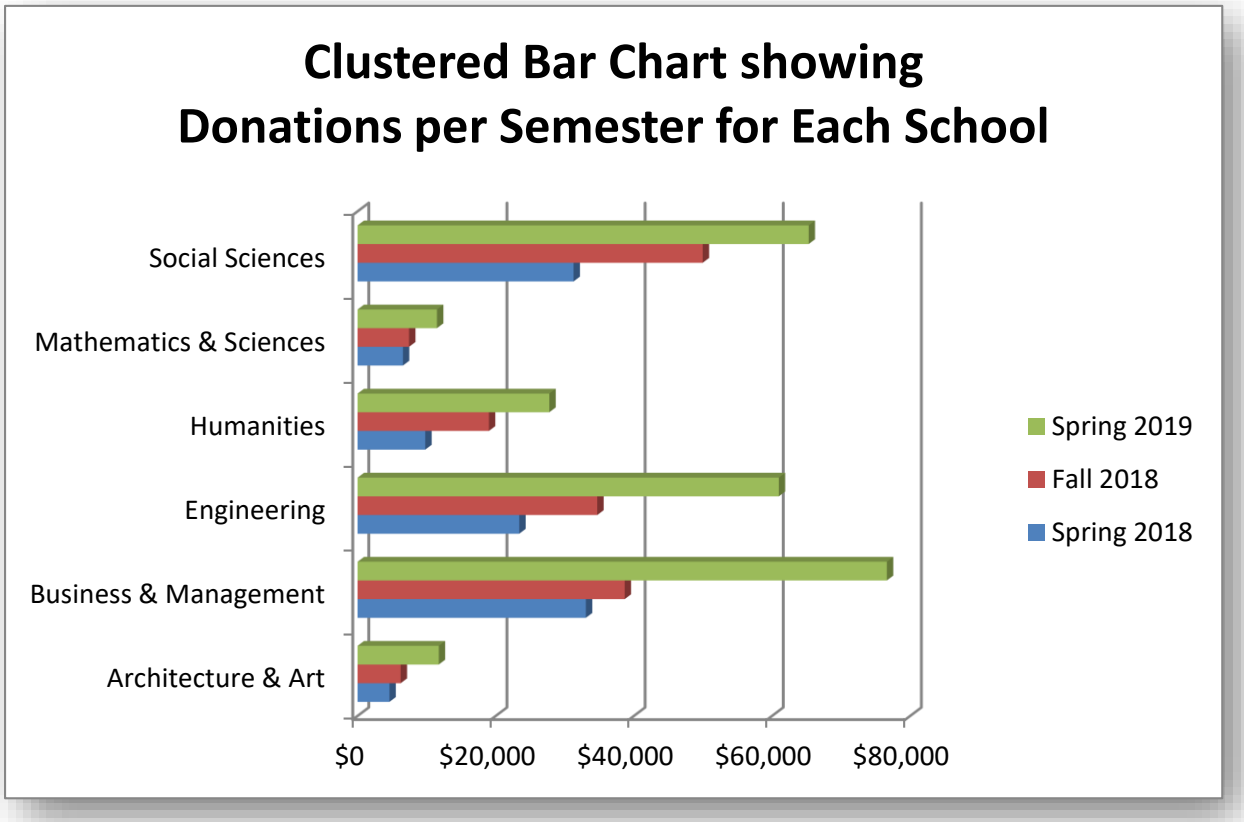
So far, we've been looking at charts for a single variable corresponding to each of several categories. However, we often have several variables for each value of the category; for example, Figure 4-12 shows the total charitable donations collected by each school in each of three semesters

Figure 4-12. Total donations by school for each of three semesters.

School	Spring 2018	Fall 2018	Spring 2019
Architecture & Art	$4,606	$6,236	$11,749
Business & Management	$33,046	$38,667	$76,623
Engineering	$23,401	$34,695	$60,979
Humanities	$9,792	$19,009	$27,754
Mathematics & Sciences	$6,560	$7,444	$11,488
Social Sciences	$31,263	$49,959	$65,300

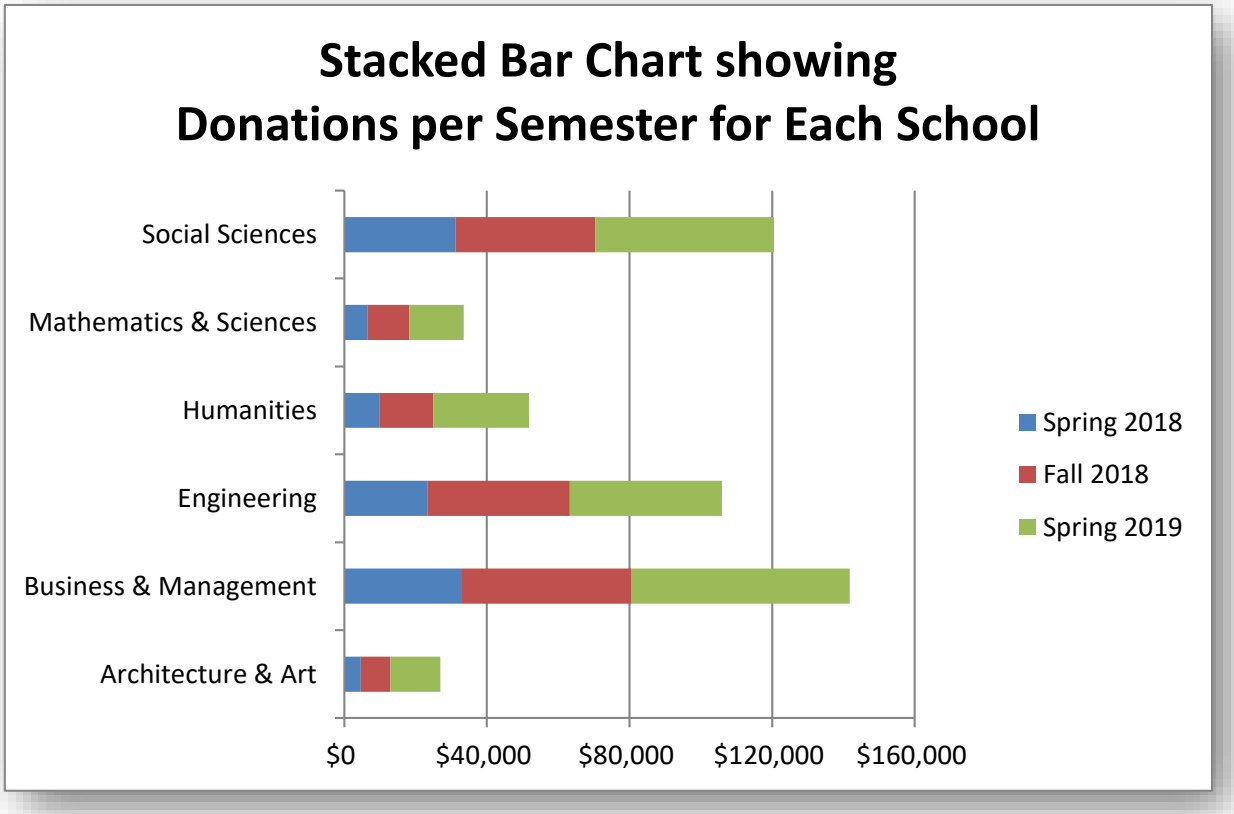
A *horizontal clustered bar chart* (Figure 4-13) can show each of the three semesters' totals as separate bars. It's easy to see the changes in donations for each of the schools over the three semesters in this display. It's also easy to compare the donations for each of the semesters for the six schools, as if the graph were a combination of three separate bar graphs overlaid to create a single picture.

Figure 4-13. Clustered horizontal bar chart.



An alternative for these data is to put the bars on top of each other in a *stacked bar chart*. Figure 4-14 shows the same data as in Figure 4-13, but it's much easier to evaluate the total donations for all three years lumped together for each school. In addition, it's also easier to see the relative size of the yearly donations within one school's record.

Figure 4-14. Stacked bar chart.

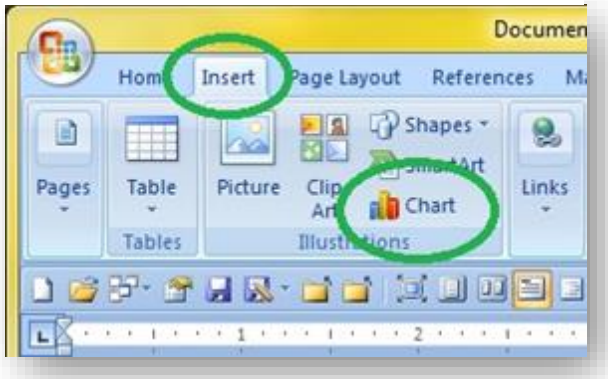


As you can see, the different layouts affect the perception and focus of the viewer. Take the time to decide which kind of format makes the most sense for your purposes. You may even want to represent the same data in various ways as your discussion moves from one issue to another when presenting your findings about a phenomenon. But an important insight is that there's no rigid rule that tells you "always use this" or "always use that;" you have to *think about your purpose* when you decide how to graph your data.

4.4 Creating Charts in WORD

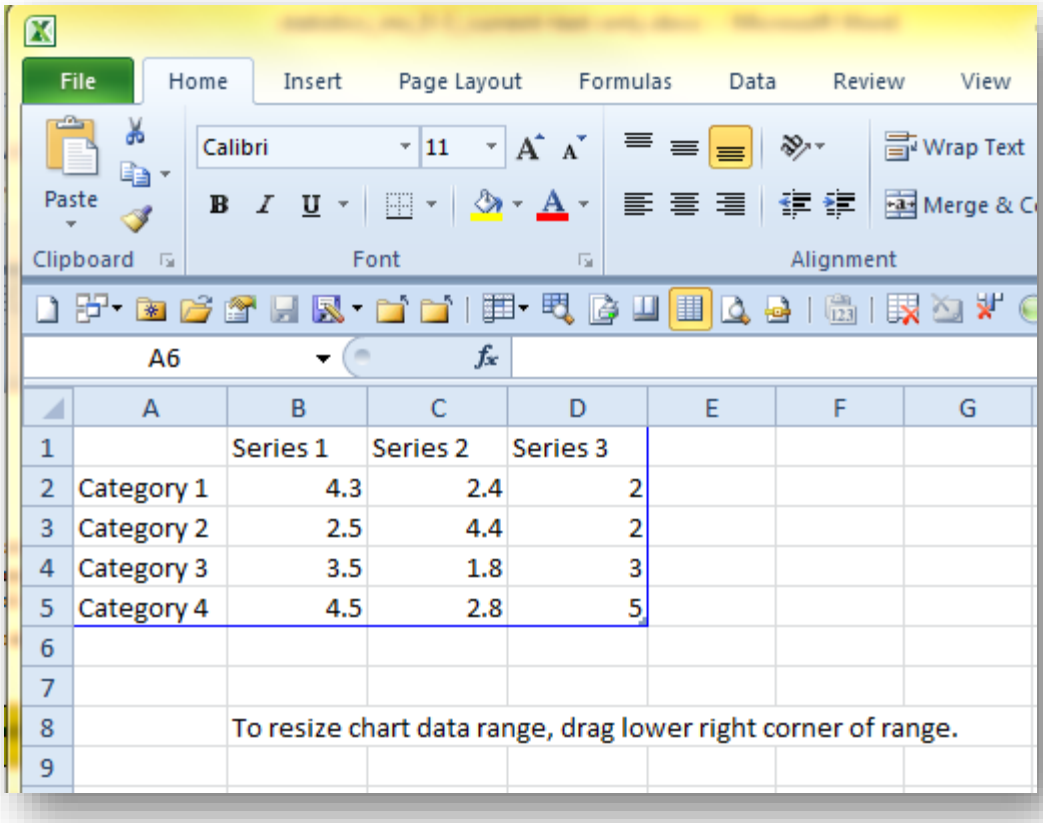
It is possible to create a chart in WORD, but the menu functions (Figure 4-15) immediately transfer control to EXCEL using a dummy table that you have to modify to suit your needs (Figure 4-16).

Figure 4-15. Word *Insert Chart* menu functions.



It makes much more sense to start your own EXCEL process, enter your data, and use the EXCEL chart tools to create your graph. You can then copy and paste the graph in whatever form (Enhanced metafile, etc.) suits your needs. For simplicity, use the Top and Bottom option in the Wrap Text options and be aware that sometimes you have to apply that option to the Figure or Table caption as well.

Figure 4-16. Dummy Excel table instantiated by Word *Insert Chart* function.



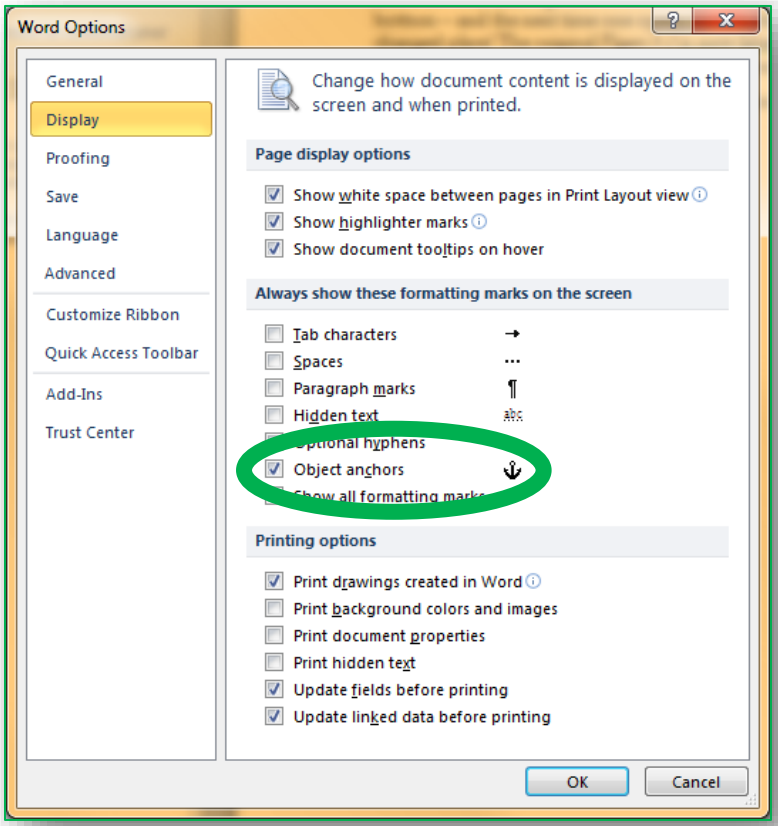
	A	B	C	D
1		Series 1	Series 2	Series 3
2	Category 1	4.3	2.4	2
3	Category 2	2.5	4.4	2
4	Category 3	3.5	1.8	3
5	Category 4	4.5	2.8	5

4.5 Managing Figure & Table Numbers in WORD

One of the most annoying features of pasting graphics into a WORD document is that their identifying numbers sometimes go out of order. One can have a *Figure 3-2* at the top of a page and *Figure 3-3* at the bottom – and the next time one opens the file, the figures are still in the same place but their labels have changed place! The original *Figure 3-2* is now labeled *Figure 3-3* and has the wrong description; the original *Figure 3-3* is now labeled *Figure 3-2* and has the wrong description. Cross-references become scrambled, too.

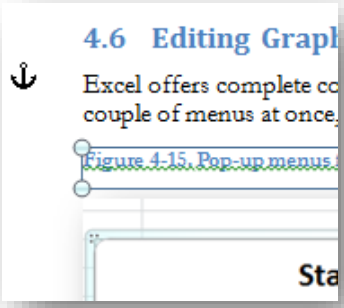
The solution is to go to the File | Options menu and check the Object anchors box, as shown in Figure 4-17.

Figure 4-17. Options for showing object anchors.



If labels, figures or tables are out of order, click on the top of the label or of the image to make WORD show you where the anchor is located (Figure 4-18), then drag the anchor to the appropriate position (paragraph) in the text ensure that the numbers are in the right order. This process also helps in positioning tables that seem eager to escape onto the next page or who jump on top of another figure.

Figure 4-18. Anchor point visible.



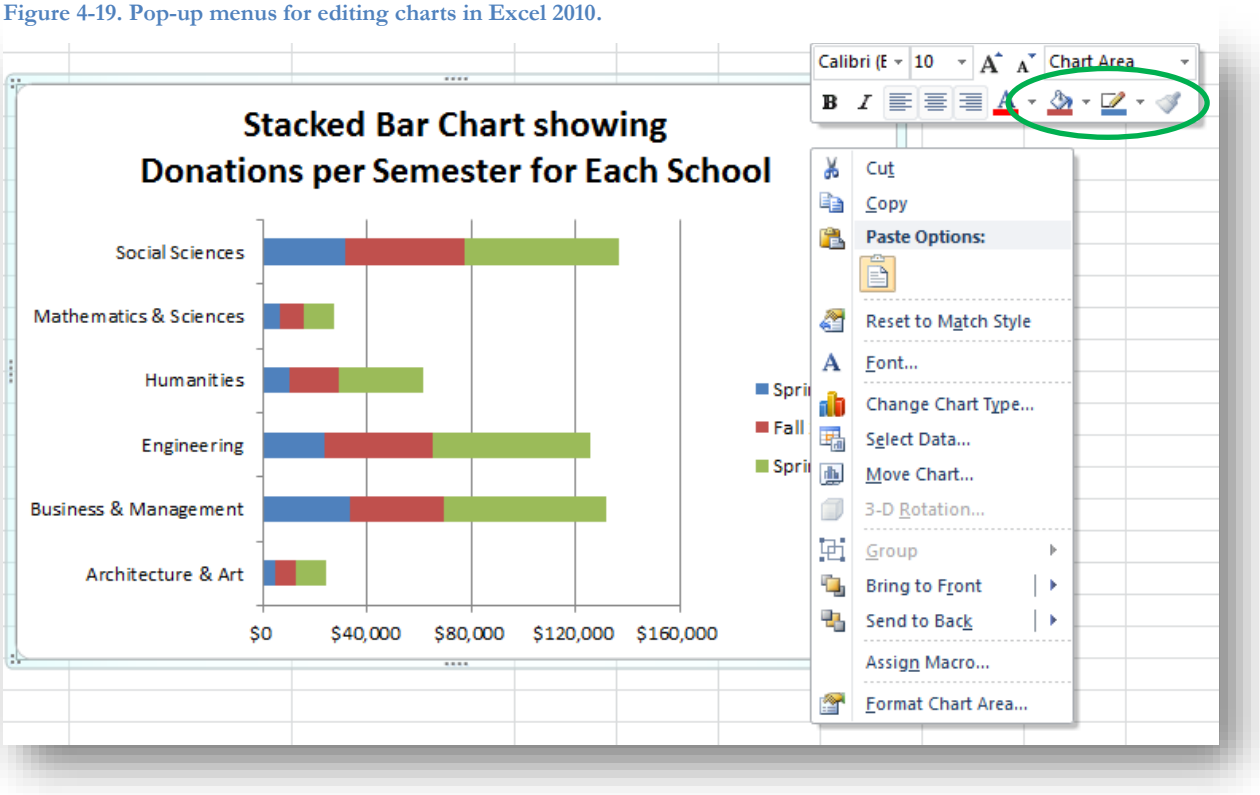
If all else fails, you can cut an image from your document, paste it into a temporary blank document, remove the caption, and then start again. You have to remove inline cross-references when you do this and insert them again too to avoid the "Error!" message in your text.

Finally, a useful trick for updating all cross-references is to highlight the entire document (Ctrl-A) and then press F9. Using the Print Preview function also works.

4.6 Editing Graphics in EXCEL

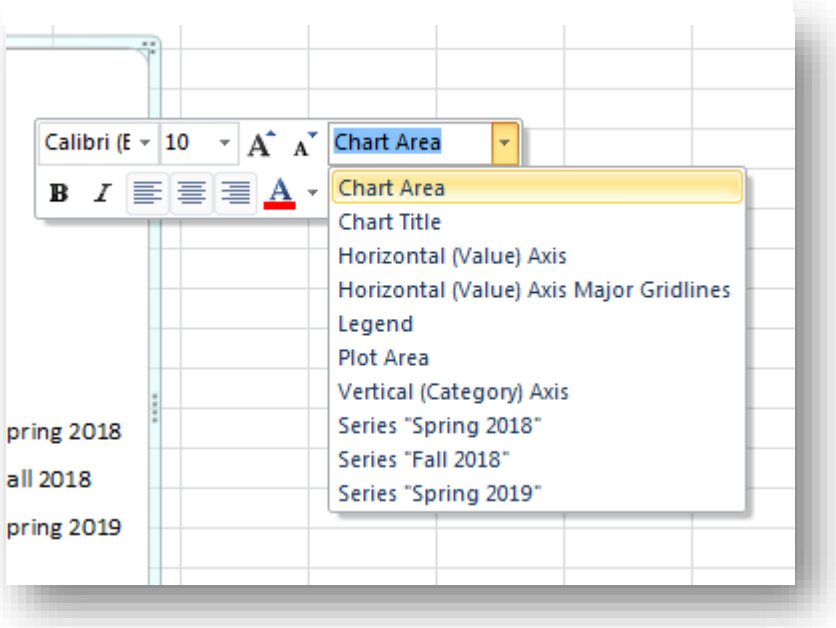
EXCEL offers complete control over every element of graphs. Right-clicking anywhere on a chart brings up a couple of menus at once, as shown in Figure 4-19.

Figure 4-19. Pop-up menus for editing charts in Excel 2010.



The pull-down menu expands as shown in Figure 4-20.

Figure 4-20. Pull-down menu for editing charts in Excel 2010.



4.7 Frequency Distributions

We often count the occurrences or frequencies of specific values in our observations. For example, we might want to summarize information about customer satisfaction shown in the following data set in Figure 4-21 (only the top and bottom of the data table are shown in this figure).

The frequency distribution in Figure 4-22 lists the number of observations that are equal to or *smaller* than the value in the row label (e.g., 0, 10, 20…) and greater than the value of the next-lower row label (except in the first row). Thus there are no observations of customer satisfaction less than 40; there are 6 entries with values between 41 and 50 inclusive, 26 entries with values between 51 and 60 inclusive, and so on. There are 5 customer-satisfaction values between 91 and 100.

An important guideline when creating histograms is that categories with fewer than 5 entries can distort statistical calculations. To be clear, **there's nothing wrong with reporting the exact counts** in the categories, but under some circumstances you may need to combine adjacent categories to reach the minimum frequency of 5. **If there are such categories, you *can* combine adjacent categories until you reach a minimum of 5.** For example, in Figure 4-22, if there had hypothetically been 3 observations in the 21-30 category and 2 in the 31-40 category, we could have defined a *21-40* category labeled *40* with 5 observations in all.

Figure 4-21. Data set with 200 customer-satisfaction data.

Customer Satisfaction (0-100)
59
78
69
81
72
75
71
81
66
81
⋮
70
80
75

Figure 4-22. Frequency distribution of customer-satisfaction data.

Customer Satisfaction (0-100)	*Frequency*
0	0
10	0
20	0
30	0
40	0
50	6
60	26
70	74
80	61
90	28
100	5

A bin boundary indicates the number of observations from just above the previous bin to the value of the bin.

Thus there were 26 values between 51 and 60 inclusive in the data represented in this frequency distribution.

4.8 Histograms

The graphical representation of the frequency distribution is called a *histogram* and is generally a vertical bar chart, as shown in Figure 4-23.

Figure 4-23. Histogram.

INSTANT TEST P 12

Find some interesting real frequency data and graph them using pie charts, clustered bar charts, and stacked bar charts. Go one step further and use a 100% Stacked Bar and decide what it tells you. Come up with a set of guidelines on what determines your choice among these options (What are you trying to illustrate?) and post your thoughts in the Discussion area in NUoodle for this week.

4.9 Creating Frequency Distributions and Histograms in EXCEL

Using the Data | Data Analysis |Histogram function of EXCEL, one can convert a set of data into a frequency distribution, organize the classes by rank, and create a simple chart – all automatically. Figure 4-24 shows the initial menu items for activation of the selection of statistical tools.

Figure 4-24. Initial sequence to access statistical tools in Excel 2010.

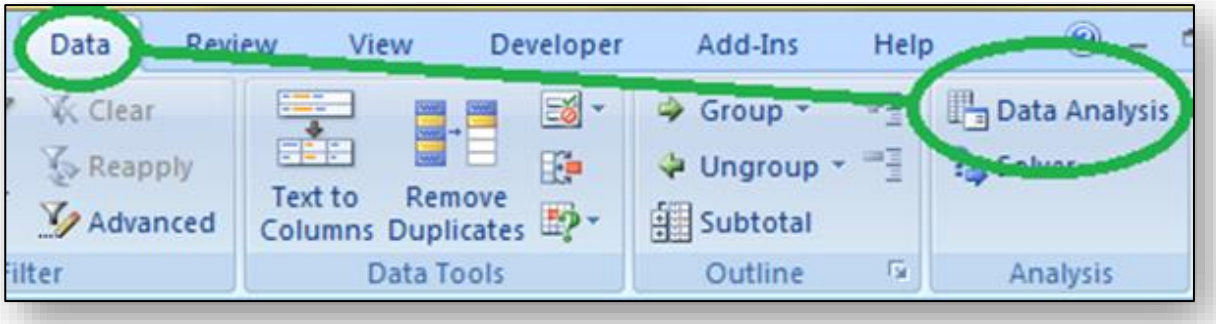


Figure 4-25. *Data Analysis* pop-up menu in Excel showing *Histogram* tool.

Clicking on the Data Analysis button shown in Figure 4-24 brings up the selection of Analysis Tools shown in Figure 4-25.

There are a total of 19 tools available; Figure 4-26 shows the remaining Analysis Tools.

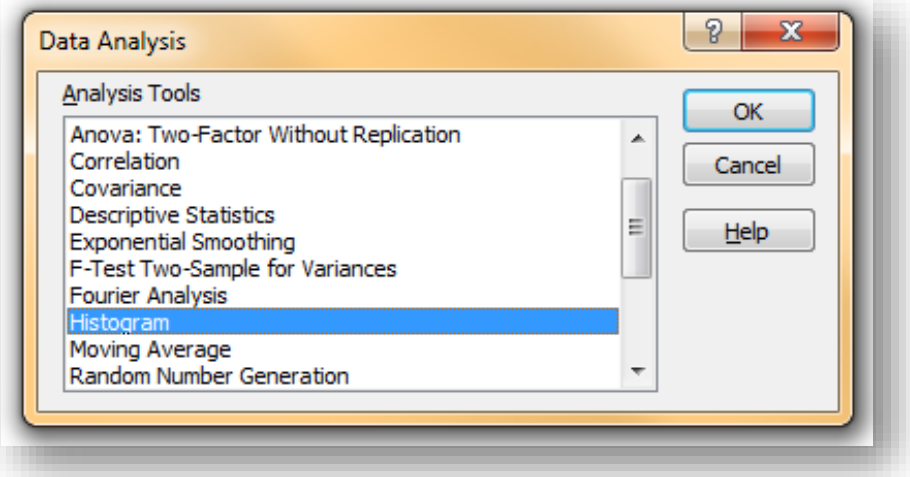
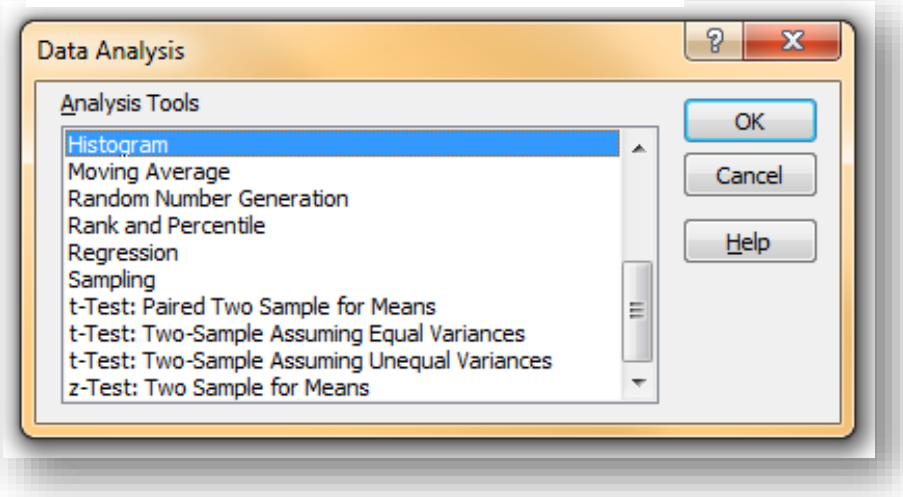
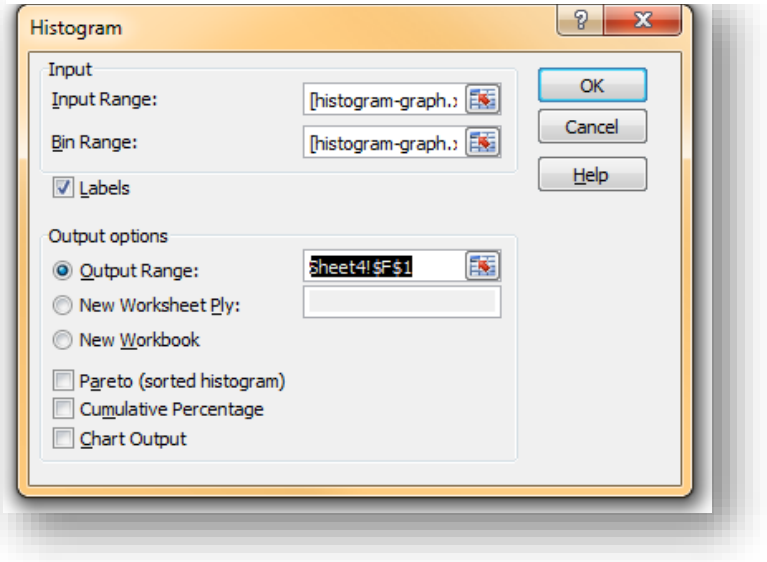


Figure 4-26. Remaining Analysis Tools.



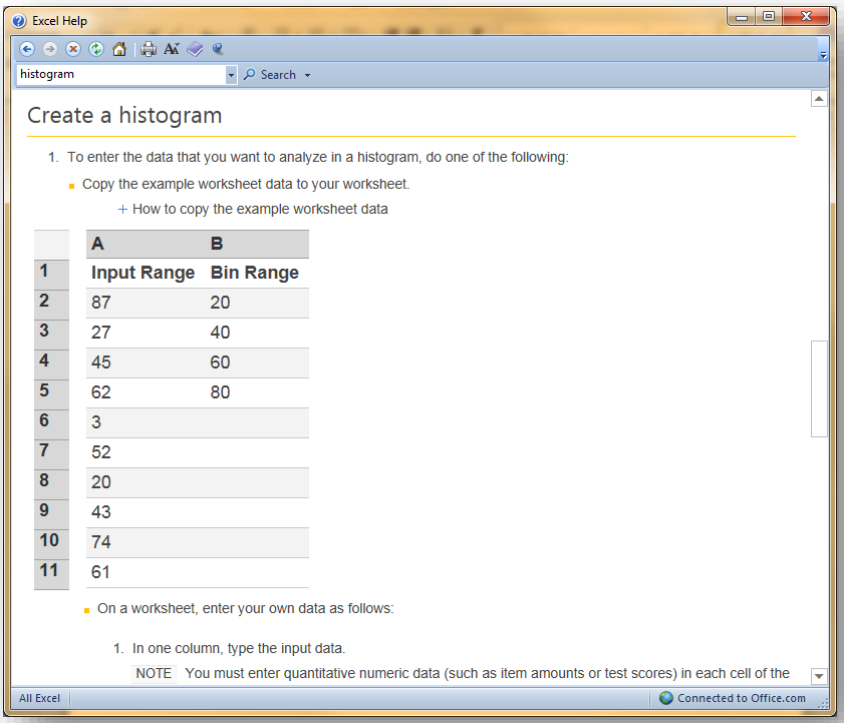
The pop-up menu shown in Figure 4-27 allows one to point to the columns – you can have multiple columns in the same chart – for input. The Bin Range stipulates where you have defined the categories you want EXCEL to tally. The Output Range (or New Worksheet Ply or New Workbook) offer options on where to put the tallies.

Figure 4-27. A *Histogram* tool pop-up menu.



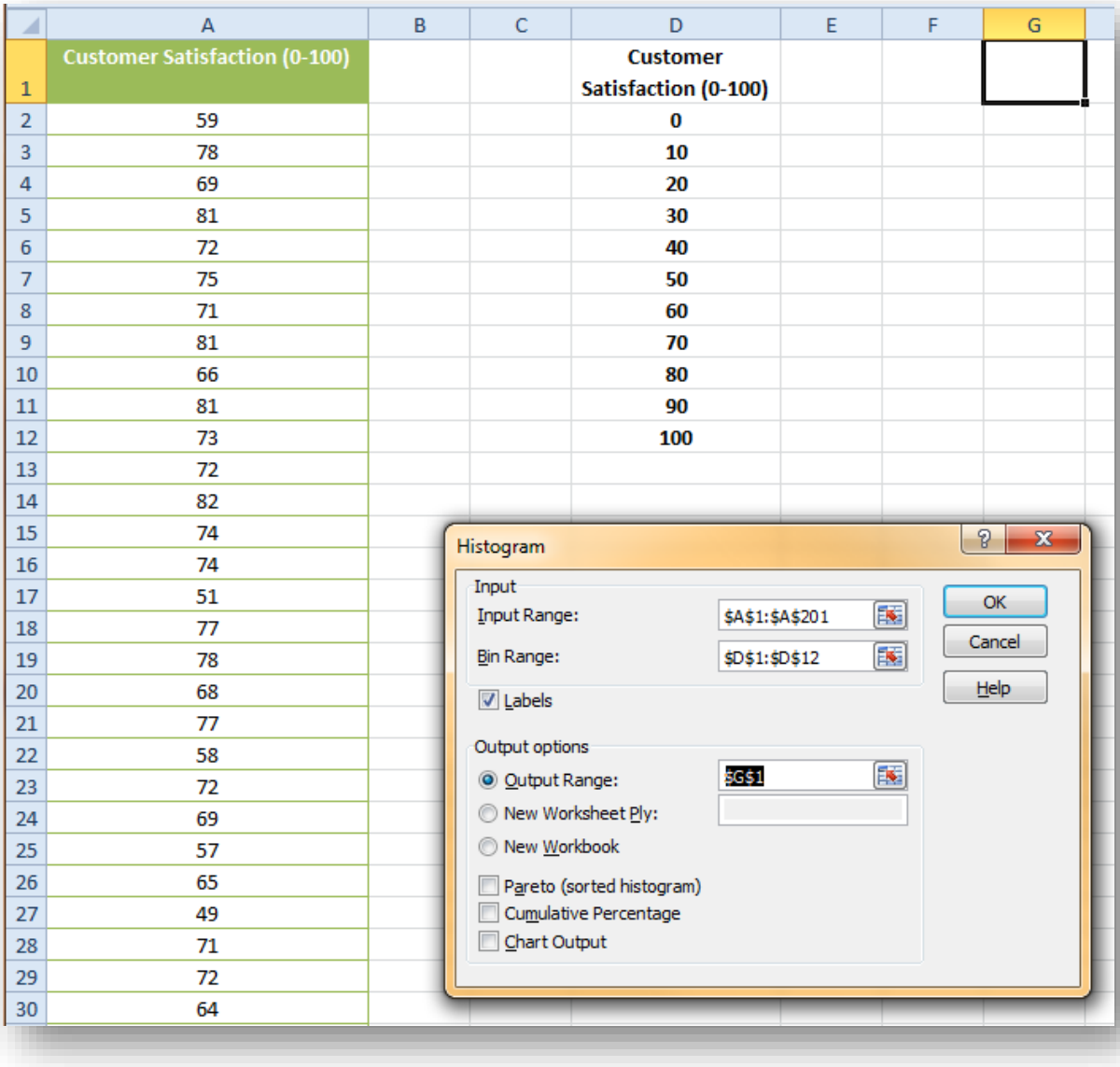
All the details of the options are available in the Help function (Figure 4-28).

Figure 4-28. Help for Histograms tool.



Be careful when enter the ranges in the menu shown in Figure 4-29. When the user clicks on the Output Range button, **a glitch in programming switches the cursor back to Input Range**. Highlighting the desired Output Range then wipes out the Input Range, to the consternation of the user and sometimes much bad language.

Figure 4-29. Defining the Input Range, Bin Range, use of Labels and Output Range for Histogram in pop-up menu.



	A	B	C	D	E	F	G
1	Customer Satisfaction (0-100)			Customer Satisfaction (0-100)			
2	59			0			
3	78			10			
4	69			20			
5	81			30			
6	72			40			
7	75			50			
8	71			60			
9	81			70			
10	66			80			
11	81			90			
12	73			100			
13	72						
14	82						
15	74						
16	74						
17	51						
18	77						
19	78						
20	68						
21	77						
22	58						
23	72						
24	69						
25	57						
26	65						
27	49						
28	71						
29	72						
30	64						

INSTANT TEST P 4-15

Practice using the Histogram function. See what happens when you click on Output Range – observe where your cursor goes. Use the tool on some data you have acquired and see what happens as you increase the number of columns. Find out what error you get if you *include* the top row of labels but *fail to check* the Labels box.

Clicking on OK generates the frequency distribution (EXCEL calls it the histogram) in the desired location as shown in Figure 4-31. The counts show the number of cases greater than the lower Bin and *up to and including* the Bin value, as shown in Figure 4-31.

Figure 4-30.Table generated by *Histogram* Analysis Tool.

G	H
Customer Satisfaction (0-100)	*Frequency*
0	0
10	0
20	0
30	0
40	0
50	6
60	26
70	74
80	61
90	28
100	5
More	0

Figure 4-31. Demonstrating definition of bins.

	A	B	C
1	DATA	*Bin*	*Frequency*
2	2.0	0	0
3	2.1	1	0
4	4.0	2	1
5	4.0	3	1
6	4.5	4	2
7	5.0	5	2
8	6.0	6	1
9	6.2	7	2
10	7.0	8	1
11	8.0	9	1
12	9.0	10	1
13	10.0	More	0

The options available for the Histogram function (Figure 4-29) are as follows:

- The Input Range and Bin Range are where one indicates the data areas. The data may be arranged in a single column or in a single row.
- Bin Range refers to a list of categories defined by the user; in this example, 0 to 100 in divisions of 10 seemed appropriate. The Bin Range must be arranged in a single column.
- Checking the Labels box allows the ranges to include descriptive labels.
- The Output Range can point to the same worksheet, as in this example, or it can be created in a new worksheet in the same workbook file (.XLSX) – or even into a new workbook.
- The option for Pareto (sorted histogram) generates additional columns showing the bins sorted downward by frequency (*modal* – most frequent – bin at the top).
- Cumulative Percentage computes the cumulative relative frequency distribution (discussed below).
- Chart Output produces a simple vertical histogram showing the frequency distribution and its graph (called an ogive, also discussed below).

INSTANT TEST P 4-16

Experiment with the options at the bottom of the Histogram menu. See what happens when you click on the various choices. Put each new histogram in a new worksheet play.

Experiment with using the New Workbook option as well.

4.10 Choosing a Reasonable Number of X-axis Values

When creating tables and charts with a numerical abscissa, how many classes should you use?

Sometimes the number of classes for which we have collected data becomes unwieldy. Imagine that we are studying the effects of a marketing campaign on the percentage of returned items in 1,250 retail stores? How could we reasonably present a table or a graph showing each individual store's results in the study? Even if we used multiple columns, such a massive table could take up several pages and would result in mind-numbing detail for our readers. Similarly, a graph with 1,250 categories on the abscissa would take up several arm-lengths of space if the names of each store were to be included.

A solution is *grouping*. We could classify the stores according to some reasonable criterion[53] such as geographical location (if all the stores are in the USA, perhaps state would be a reasonable basis for grouping, or maybe areas such as "northeast, southeast, north central, south central, northwest, and southwest).

Alternatively, an analyst might want to focus on the size of the stores and group the results by the total gross revenues; e.g., $1M to $4.9M, $5M to $9.9M, and so on.[54]

Some guidelines for grouping data for statistical representation and analysis:

- In general, *somewhere between 10 and 30 groups* seems reasonable for tabular and graphical presentations.
- If your grouping criterion is numerical (e.g., total revenue, number of employees, number of shares sold, level of production errors) you can estimate the optimal interval size by calculating the range (the largest value minus the smallest value) and then dividing the range by 10 and also by 30.
- For example, if you have share prices as the criterion for grouping companies in an analysis, and the cheapest share costs $138 whereas the most expensive share costs $3,848, then the range would be $3848-$138 = $3710.
- So the smallest interval (the one producing 30 groups) would be $3710/30 = 123.7 or about 124.
- The largest interval (the one producing 10 groups) would be $3710/10 = $371.
- Avoid peculiar groups such as 27.3:37.2; groups starting and ending on 0s and 5s are common (e.g., 100, 105, 110… or 1200, 1400, 1600…). You don't have to start at the minimum, either; if your minimum were 128 and your groups were 20 units wide, you could start with the first group at 120:139, the second at 140:160 and so on.
- You might pick something like an interval of $200 or $250 to keep things neat; that would generate $3710/$200 = 18.5 groups which means 19 groups in all (you can't have a fractional group). The $250 interval would produce 15 groups.
- The first $200-wide group in this example could thus start off being $0 to 199; the second group by share price would be $200:$399; the third, $400:$599. The last group under this scheme would be $3800:$3999.
- Avoid categories with fewer than five observations. Combine adjacent categories if necessary to arrive at a minimum of 5 occurrences per category.
    - You may want to sort your initial groups by the number of data and then combine adjacent low-frequency groups until you reach the minimum of five observations per group.
    - For example, if there were only two entries in the $0:$199 group described above and six in the $200:$399 group, you could combine those two into a $0:$399 group and have eight in it.

[53] A *criterion* is a basis for judgement . The plural is *criteria*. Don't say or write "the criteria is" or "the criterion are." From Greek κριτεριον, *kriterion* from κριτες, *krites* = judge. Our words *critic* and *criticism* come from the same root.

[54] Groups are usually indicated using colons ($0:$199) or ellipses ($0...$199) but not dashes ($0-$199). We avoid hyphens and dashes to avoid ambiguity with subtraction when negative numbers are included in the ranges (e.g., -15 - -10, which looks like a mistake).

4.11 Problems with Disparate Quantities

Suppose the Urgonian Corporation carried out a survey of customer satisfaction in seven markets around the solar system. Figure 4-32 shows the results in terms of raw data and of percentage of positive responses. The layout of the chart is *vertical bar* only to save space on the page: usually it would be horizontal bars.

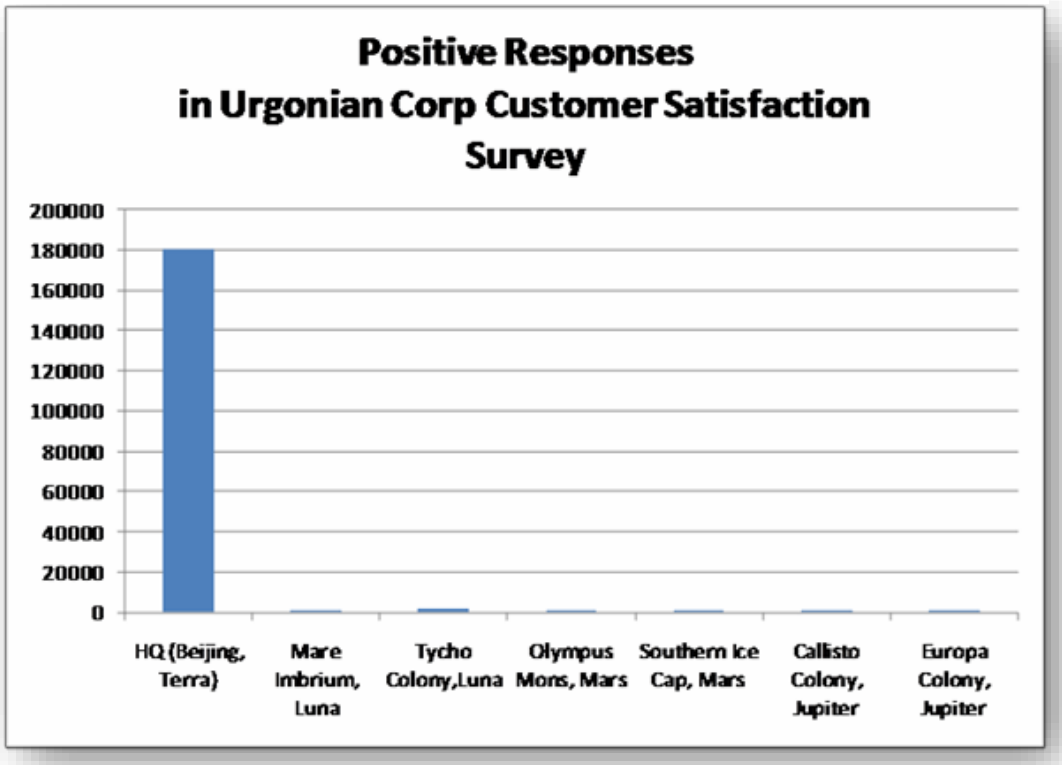
Figure 4-32. Urgonian Corporation survey results.

SURVEYS 2218.04.22	Returned	Positive	% Positive
HQ (Beijing, Terra)	324,932	180554	56%
Mare Imbrium, Luna	1847	1231	67%
Tycho Colony, Luna	2950	1915	65%
Olympus Mons, Mars	372	278	75%
Southern Ice Cap, Mars	414	266	64%
Calisto Colony, Jupiter	112	77	69%
Europa Colony, Jupiter	185	137	74%

If we try to represent the raw data about the *number* of positive responses in a histogram (Figure 4-33), we immediately run into trouble: the range of total responses is so large (from 180,554 down to 77) that it is impossible to show anything meaningful about the lower frequencies on the same graph. All we can tell is that there were a lot of responses from Beijing; details of all the other sites are obscured because they are so tiny compared with the total from Beijing.

In general, it is not useful to try to show wildly different quantities on the same graph using a linear scale for the ordinate if the maximum is more than about 20 times the minimum. In the data for Figure 4-33, the largest value is *2,345 times larger* than the smallest value.

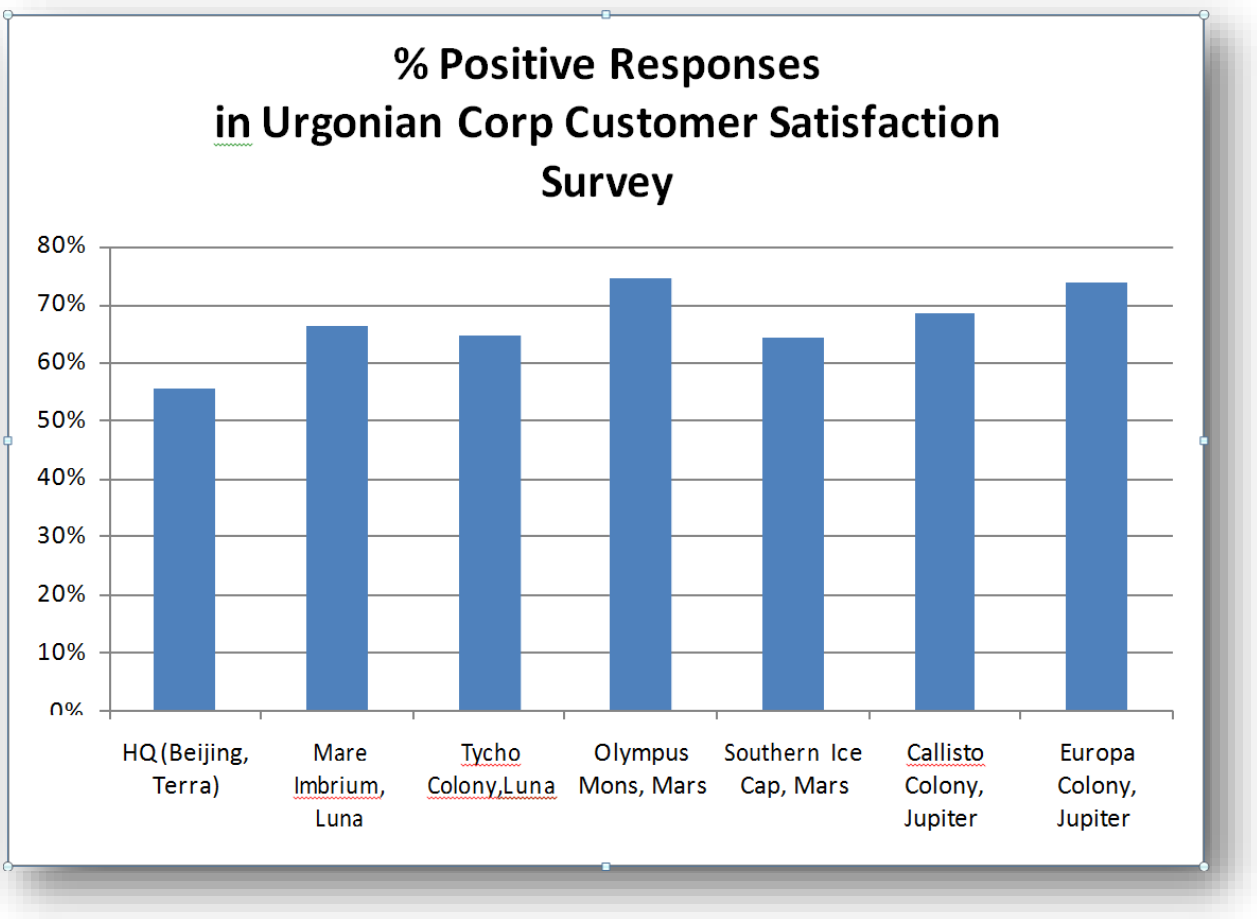
Figure 4-33. Useless vertical bar chart with disparate values.



Second, showing only the positive returns tells us nothing about the relative *proportion* of positive returns, which are more likely to be interesting. We must always think about whether a representation is meaningful before taking our time to create it.

Figure 4-34 shows the percentages of positive returns for the Urgonian Corporation customer satisfaction survey. Since those percentages vary from 56% to 75% and the theoretical limits are 0% and 100%, we should be OK.

Figure 4-34. Vertical bar chart of percentages of positive responses.



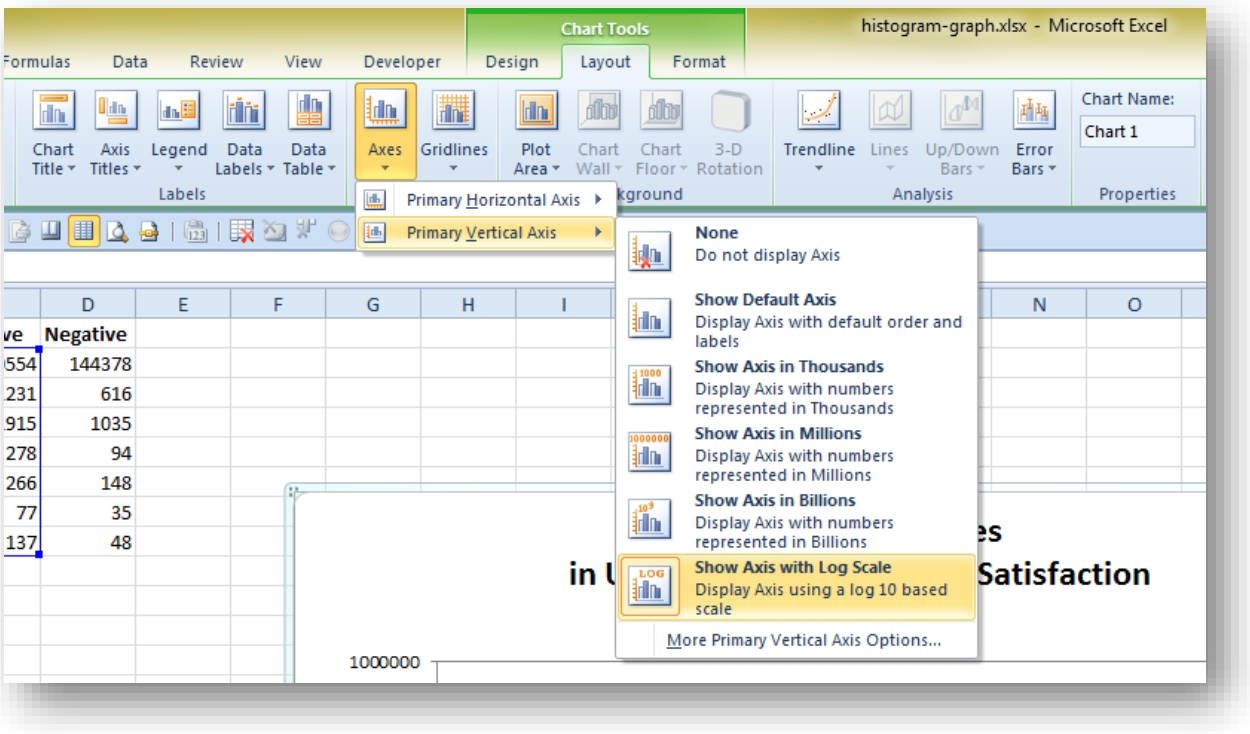
INSTANT TEST P 4-19

Experiment with the data in Figure 4-32 to create horizontal bar charts with these values. Try out different options such as the shape of the bars (3D, cones…) and examine the effects of options such as those for the Y-axis scale.

4.12 Logarithmic Scale on the Ordinate

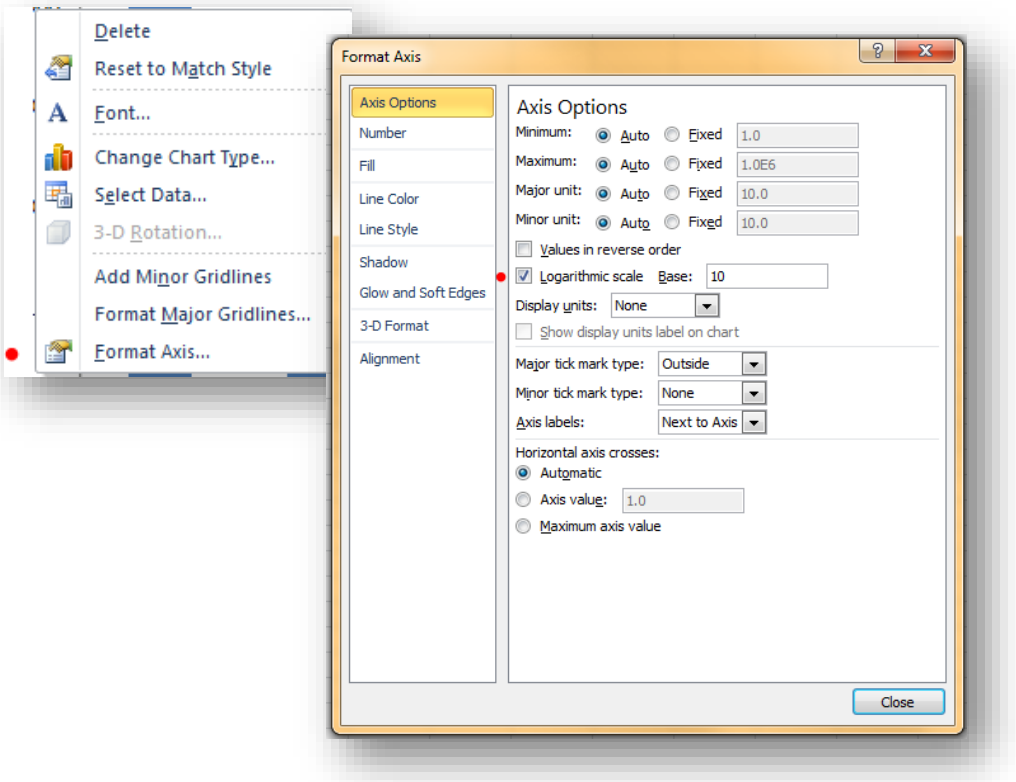
If you still need to represent the original disparate data on the Y-axis, you can use the logarithmic-scale option shown in Figure 4-35.

Figure 4-35. Excel *Chart Tools* menus to define logarithmic scale on ordinate.



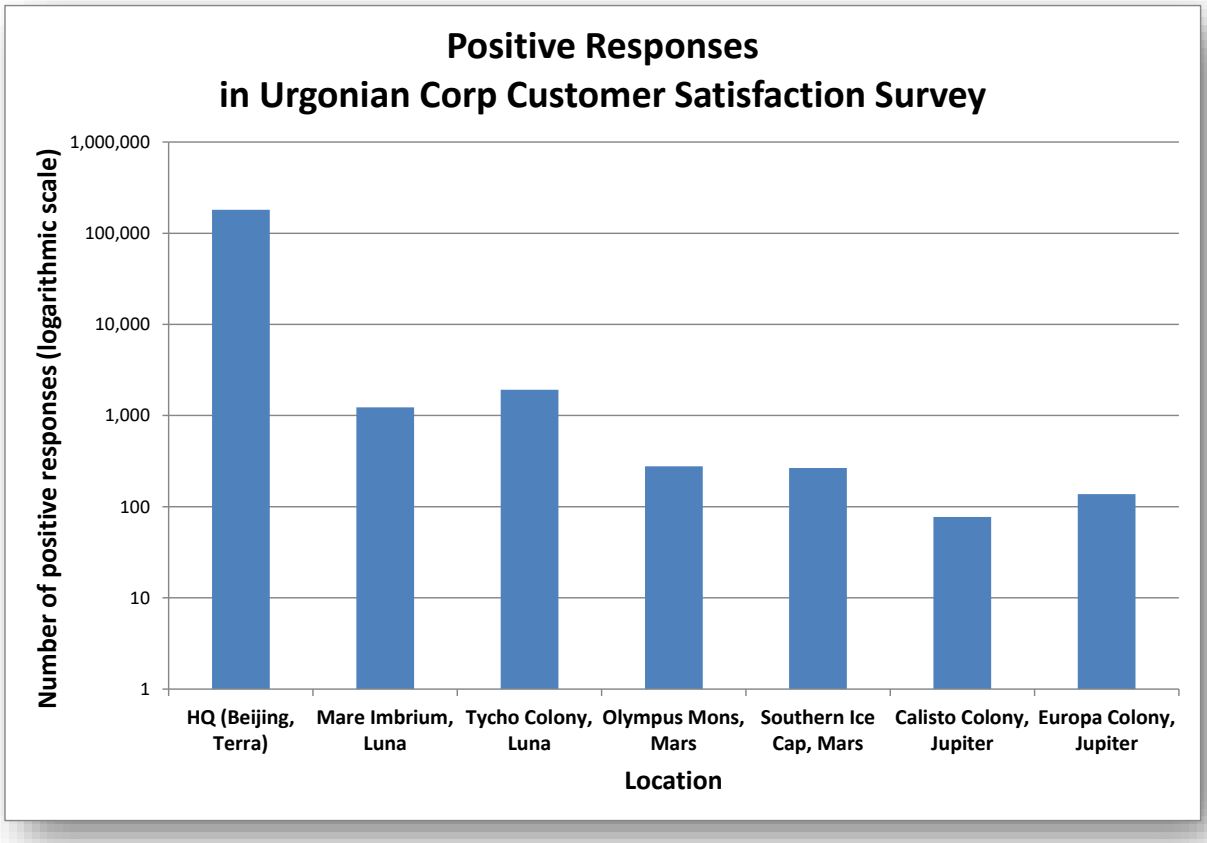
An alternative is to right-click on the ordinate scale to bring up the pop-up menu and select the Format Axis… function to bring up the Format Axis pop-up menu as shown in Figure 4-36.

Figure 4-36. Right-click menu for vertical axis and *Format Axis* pop-up.



The results are shown in Figure 4-37. Notice that the ordinate (Y-axis) is correctly identified as using a logarithmic scale.

Figure 4-37. Urgonian data with log scale Y-axis.



A word of warning: log scales make perfect sense to people who are familiar with them but can be confusing for people unfamiliar with the concept of logarithms.

- For example, a value that is twice as large as another on a $\log_{10}$ scale (as in the example here) is ten times larger in reality.
- The vertical bar for the number of positive responses from the Mare Imbrium site is half the size of the bar for the data from HQ (Beijing) – but there are 100 times more positive responses in Beijing than in Mare Imbrium.

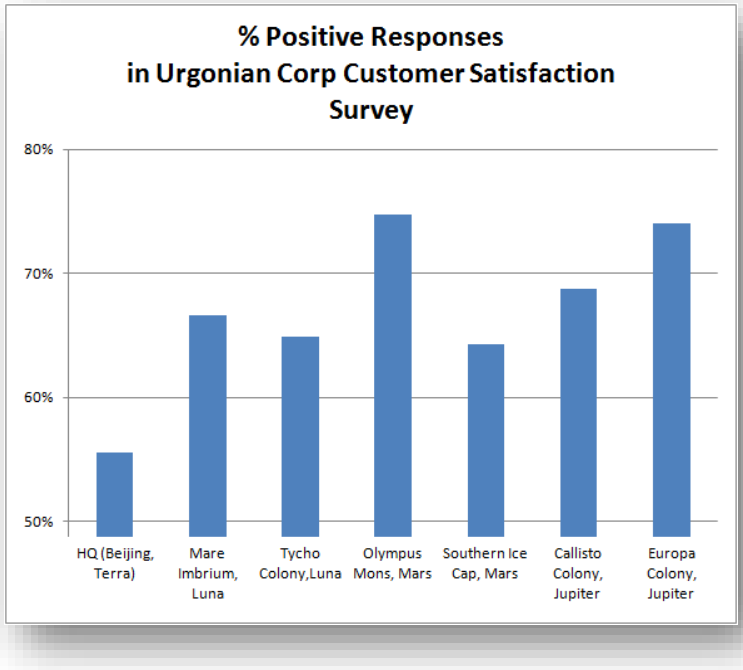
Some additional recommendations:

- If you use such a graph with naïve users – people unfamiliar with the very concept of a logarithm – there is a risk that they will misinterpret the data.
- On a nastier note, beware anyone who uses a log scale without explicitly mentioning it; they may be trying to trick the user of the graph into a misunderstanding.
- Finally, don't fall into the error of labeling the Y-axis as "Log of [whatever it is]." If we were to label the values with the actual logarithms in Figure 4-37, the Y-axis labels would read (from the bottom) 0, 1, 2, 3 and so on.

4.13 Truncating the Ordinate

One of the most serious errors that beginners (or dishonest professionals) can make is to inflate the apparent differences among groups in a histogram by cutting off the bottom. For example, the distortion in Figure 4-38 cuts off the bottom 50% of the histogram ordinate and grossly distorts the differences among the divisions of the corporation. A casual viewer might think that the Olympus Mons division has a positive rating several times that of the Beijing division, but in fact the difference is not nearly as large as the impression created by the dishonest representation.

Figure 4-38. Survey results with truncated Y-axis.



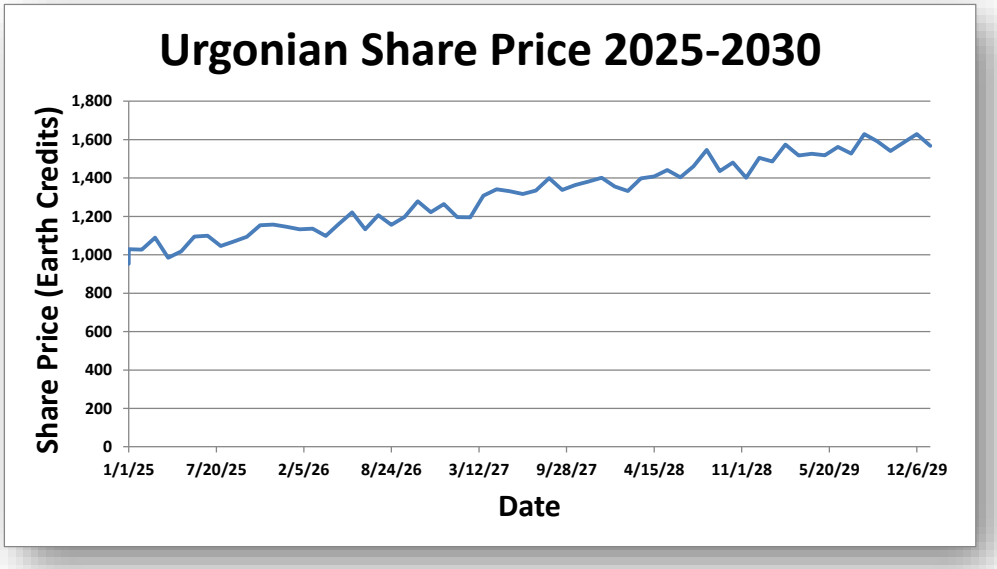
In general, *do not set the intersection of the abscissa and the ordinate (the origin) anywhere but zero* when creating a histogram. Even if you are graphing some other relation not involving frequencies (e.g., price fluctuations in stocks), be wary of arbitrarily cutting off the bottom of the graph and thus potentially misleading viewers into magnifying differences.

INSTANT TEST P 4-23

Experiment with the data in Figure 4-32 to create a *horizontal* bar chart with these values. Change the scale of your new horizontal scale (% responses) to start at 55% and comment on the effect in your chart. Go back to starting the horizontal scale at 0 and compare the impressions created by the two graphs. Why is one better than the other in this case? What would you say if someone told you that there had never been a value lower than 55% in this statistic – would that change your thinking? Why?

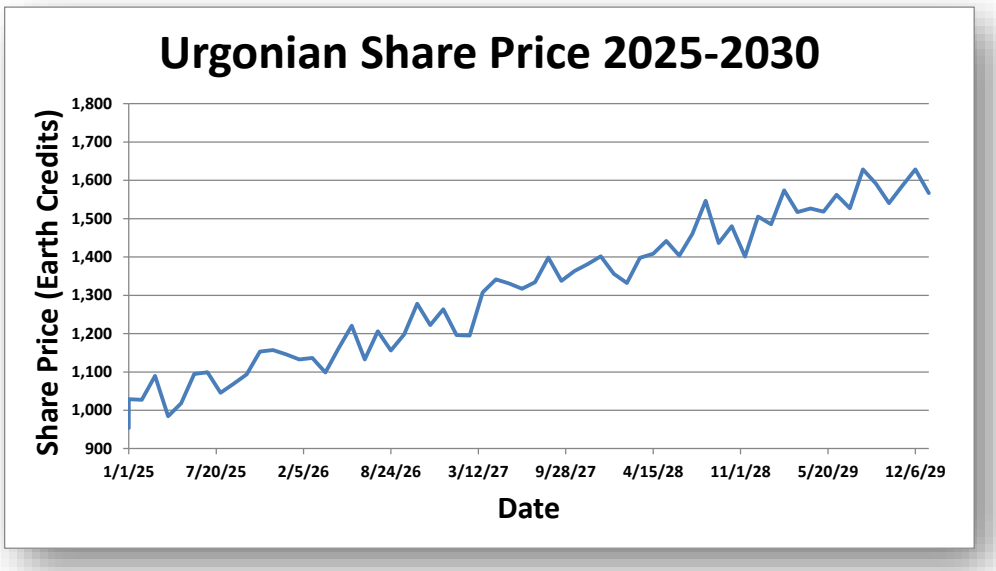
On the other hand, in some applications, it is accepted (despite protests from statisticians) that the origin should reflect the nature of the fluctuations. Figure 4-39 shows prices for a corporation over a five year period. Even without regression analysis or curve fitting, it is obvious that the price has been rising modestly over the period shown.

Figure 4-39. Time series for share prices using 0 as start of ordinate.



However, such time series typically start at some level that reflects the consistent lack of data falling below some arbitrary level. If no one has seen the share price below, say, 900 credits, many financial analysts and journalists will truncate the ordinate as shown in Figure 4-40. Comparing the two graphs, one can see that both the difference between later and earlier prices and the rate of growth in share price seem to be much greater in the graph with the truncated ordinate. This impression might be helpful to dishonest brokers but it could be dangerous for naïve stock traders.

Figure 4-40. Time series for share prices using 900 as start of truncated ordinate (DANGEROUS).



4.14 Selecting Non-Random Sections of a Data Series

One of the most common occurrences of misleading graphs is in published articles where the authors or editors have an axe to grind. By adjusting the origin of the graph and selecting a subset of the available data, they can create whichever impression they wish. For example, examine the carefully crafted graph in Figure 4-41.

Figure 4-41. Misleading graph using selected data and distorted Y-axis (HORRIBLE).

This image is based on the last five stock prices used to create the graph in Figure 4-39 and Figure 4-40 and could be part of an article deliberately trying to give the wrong impression that the share price "has been erratic for some time and is now falling steeply." Leaving out all the previous data and focusing only on this extract – coupled with setting the Y-axis to give a grossly exaggerated view of the changes – can give whatever impression the writer wants to convey. Whenever you see a grossly truncated ordinate or a small number of data points you should get your skepticism antennae vibrating and look for possible trickery. **And don't create this kind of monstrosity in your own work!**

INSTANT TEST P 4-25

Find a real time series of data for something you are interested in using resources on the Web or in journals. Prepare a graph for an extended time period. Then take the graph and cut the period to a tiny fraction of the data available, selecting a section that gives the opposite impression of what you see in the overall graph. Show the image of the new image on screen (or in a printout) to some friends and ask them to summarize their impressions about the phenomenon they see displayed in the chart. Then show them the graph for the extended period and ask them what they think about that. Keep a record of the answers and prepare table showing, for each person, their comments on the fragment and their comments on the whole graph. Summarize your findings and post your conclusions in this week's NUoodle discussion group.