Appendix IV: Creating and Interpreting Graphs

Big Idea #1: A good graph tells a story about the relationship between two (or more) things.

We present data in graphs in order to identify trends and make predictions. Translating between graphs and words is an essential skill for success in science (as well as business and other fields). So let's start practicing.

Consider the graph below. What do you see? What story can you construct to explain the graph? What additional information might be helpful in order to fully understand the graph? For instance, you might be wondering, "What is the Case-Shiller '10 city composite' index?" "What exactly do the dollar values on the x-axis represent?"

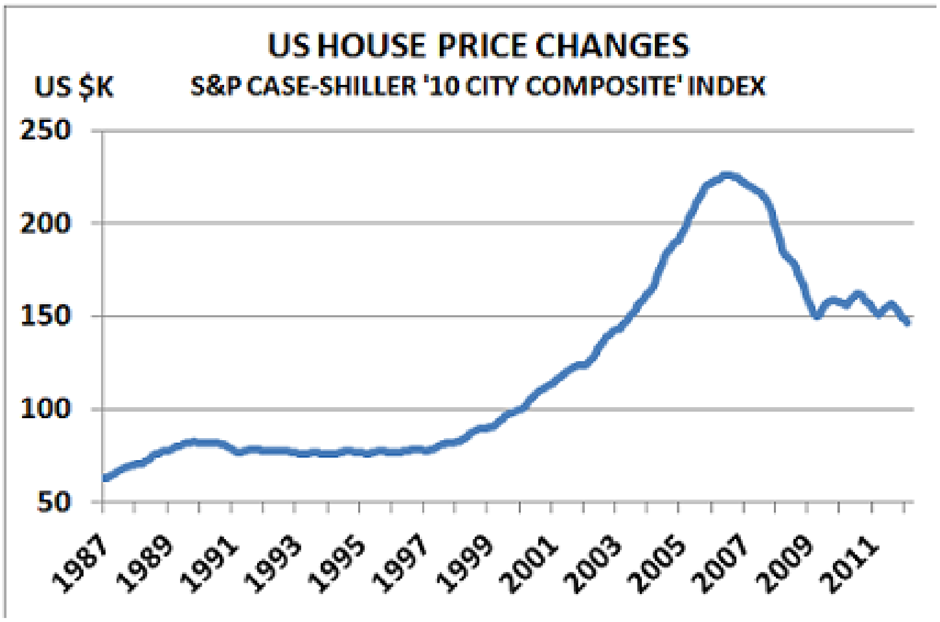

Figure 1: The Case Shiller 10-City Composite Index measures the average selling price of homes sold in ten of the largest metropolitan markets in the United States. For more information on how it is calculated, see http://www.standardandpoors.com/indices/sp-case-shiller-home-price-indices/en/us/?indexId=spusa-cashpidff--p-us----.

Use the graph to answer the following questions. A discussion of the answers is at the end of the appendix.

1. What are the two variables being compared on this graph?
2. Estimate the average rate of increase of the index (in thousands of dollars per year) from 1987 to 1997.
3. Estimate the average rate of increase of the index (in thousands of dollars per year) from 1997 to 2006.
4. By what percentage did the index decrease between 2006 and 2009?
5. How does the graph relate to your understanding of current events?
6. Estimate what the value of the index will be in 2015. Explain your reasoning.
7. Estimate what the value of the index will be in 2040. Explain your reasoning.
8. Estimate what the value of the index was in 1960. Explain your reasoning.

Big idea #2: When you make a graph, you should include enough information for the viewer to "read" its story clearly.

Every graph you make in this course should include the following:

1. A descriptive title, preferably something that gives more information than just stating the variables being compared.

2. A consistent scale on each axis. In other words, a set distance on the x-axis should always be worth the same amount. You should not make the first inch of the axis worth 10 units and the next inch worth 50 units. That kind of inconsistency distorts the story of the graph. Note that it is ok to have different scales on the x and y-axes. So one inch on the x-axis might represent a different number of units than one inch on the y-axis.
3. Labels for each axis, including units.
4. If your graph requires additional explanation, you should include it in a caption. For instance, the graph in Figure 1 needed some explanation of the meaning of the Case Shiller Index.

Making graphs electronically – How to use a spreadsheet program:

While graphing data by hand is an important skill, in this course you will be expected to make most of your graphs on a computer. There are many programs available to do this, but I recommend using either Microsoft Excel or Google Spreadsheet. Both of these platforms are spreadsheet programs. Spreadsheets are important tools for organizing information and making calculations, and you are likely to use them extensively in the future, whether you pursue science or not.

A spreadsheet is laid out in a series of columns and rows. The rows are numbered and the columns are lettered. Each cell (individual box) can be referred to by its number and letter. For example, the box in column C row 5 is called "c5". While the details vary from program to program, the basic idea is the same. The instructions below walk you through four of the key features of a spreadsheet that you will want to use this year.

1) Performing a calculation
When you want a spreadsheet program to do a calculation, you give it a signal. That signal is an equals (=) sign. Go to any cell in the spreadsheet and type in "=2+2". Then hit "Enter." What pops up in the cell? Ok, erase that cell now.

You can also use this feature to do calculations using numbers in other cells. So try this: Go to cell a1 and type in a number. Now go to cell b1 and type "=a1+2". Then hit "Enter." What happens?

You can tell the program to find the sum of a bunch of numbers, the average, the median, etc. Go to column d and type in 5 numbers in the first 5 cells. In the sixth cell, type "=sum(". After typing the open parenthesis, use the cursor to highlight the five cells above. What should pop up in the formula bar is "d1:d5". Now close the parentheses and hit "Enter." Cell d6 should now show the sum of the 5 numbers. You can find the average by going to another cell and typing "=average(d1:d5)".

2) Repeating a calculation multiple times
This is one of the features that make a spreadsheet program so useful. Keeping your numbers in column d, go to cell e1. Type in "=d1*20". Then hit "Enter." What should pop up is the number in cell d1 times twenty. Now click the lower right-hand corner of cell e1 and drag it down across cells e2, e3, e4, and e5. Each of those cells will now contain the number to its left times twenty. The same calculation was repeated for all the numbers!

3) Making a graph of a data set
Every spreadsheet program should have some graphing capability, although the specific instructions vary widely. I have found that the easiest way to make a graph is usually to arrange the data by putting the x-

axis values in one column and the y-axis values in the column immediately to the right. Then highlight the two columns of data and select "Insert → Chart" (or some similar command). The type of graph we will usually make is usually considered a "scatter" plot. Once you have made a graph, you'll need to customize it. Make sure you give it a descriptive title, label the axes, etc. You can choose to either show gridlines or not, set the minimum and maximum values on each axis, and choose other options to make your graph tell its story as clearly as possible.

4) Fitting a data set to a function
When we graph data, we are looking for a pattern. Once we identify a pattern, we can use it to make predictions. The way we describe the pattern mathematically is with a function. The function could be linear (e.g., y = 0.024x), quadratic (e.g., $y = 2.4x^2 - 3x + 2$), exponential (e.g., $y = 1.7e^{-0.37x}$), or anything else, really. A spreadsheet program can apply mathematical algorithms to find out what function best fits our data set. However, we need to tell it what kind of function we're looking for. In this course, we will usually be looking for a linear function. We can refer to this process as finding a "line of best fit."

The instructions below are for how to insert a line of best fit in Microsoft Excel. In a different spreadsheet program, you may or may not be able to use them, but you can play around to try to figure it out.

- Click on the graph.
- In the newer version of Excel, go to the "Layout" menu and find "Trendline." In the older version, go to the "Chart" menu and find "Insert Trendline."
- Select "More Trendline Options."
- Make sure that "Linear" is selected. Check the boxes for "Display Equation on chart" and "Display R-squared value on chart."
- After you click "OK," or "Close," the line should appear on your graph, accompanied by the equation and the R^2 value.

The equation will tell you the slope and y-intercept of your line. The R^2 value is a statistical quantity that tells you how well your data fit the straight line. An R^2 value close to 1 means that the fit is very good. A lower R^2 value means that the fit is not very good. Therefore, if you are comparing several graphs, the one with the R^2 value closer to 1 is a better fit for the data.

Answers to questions:

1. The graph compares the value of the "Case Shiller 10-City Composite Index" and time.
2. The average rate of increase can be found by calculating $\frac{change\ in\ index}{change\ in\ time}$. We can estimate that the index changed by about \$10,000 (from \$70,000 to \$80,000) over a ten year period.
$$average\ rate\ of\ increase = \frac{\Delta index}{\Delta time} = \frac{\$10{,}000}{10\ years} = \$1000/year$$
3. We can estimate that the index changed by about \$150,000 (from \$80,000 to \$230,000) over a nine year period.
$$average\ rate\ of\ increase = \frac{\Delta index}{\Delta time} = \frac{\$150{,}000}{9\ years} = \$17{,}000/year$$

4. The index dropped from \$230K to \$150K, a decrease of \$80K.

$$\% \; decrease = \frac{amount \; of \; decrease}{original \; amount} \times 100\% = \frac{\$80K}{\$230K} \times 100\% = 35\%$$

5. There's not really a correct answer for this one. But you may or may not be aware that the huge growth in housing prices and, more importantly, the subsequent drop, are largely responsible for starting the "Great Recession" in which we find ourselves.

For questions 6-8, essentially, we're trying to extrapolate, or predict data that's outside our current range based on the trends we see. For each question, we need to decide what trend to use. Do we want to use a short term trend, a long term trend, something else? Our best estimates are below.

6. 2015 is only 3 years away. For such a short period of time, we would probably only look at recent trends in the data. Over the last 3 years, the index has not changed much. If anything it is down slightly, but it has bounced up and down around \$150K. Based on the data available, then, it is likely that the index will still be around that value 3 years from now. Of course, given lots more economic information, we may be able to come up with a better prediction, but based on what we're given, this is the best we could do.

7. Now we're looking 28 years into the future. For this kind of projection, we would probably want to consider the longer term trend in the graph. We could make an assumption that the graph will be essentially linear over time. In that analysis, we should think aboput what the slope of the graph has been over the long term. Over the past 25 years, the index has increased by about \$80K. Considering that trend, we might expect it to increase by another \$80K over the next 25 years. So an estimate of \$230K for 2040 seems reasonable. It's interesting to note that this is about the value of the index at the peak of the pre-recession "bubble."

   Another way to look at the trend is to consider the percent increase over time, and assume that the percent change will be roughly the same. This leads to an exponential graph. In this analysis, note that the index has essentially doubled (from \$75K to \$150K) over the last 25 years. So we might expect the index to double again between now and 2040. This would give us an estimate of \$300K for 2040. That leaves us with a wide window, saying that we expect the index to be somewhere between \$230K and \$300K in 2040. Of course, many many things can happen between now and 2040, but again, this is the best estimate we can make based on limited data.

8. In order to project backward, note that a linear analysis will NOT work. We said that over the past 25 years, the index has increased by roughly \$80K. Extending that trend backward by 25 years would have put the index at -\$10K in 1960. Clearly, this would be impossible, because it would mean that the average person was paying someone else \$10,000 to take their house. So, it would make sense to follow the exponential trend and assume that the index had doubled over the previous 25 years. This would put the value of the index at about \$30K-\$35K in 1960.