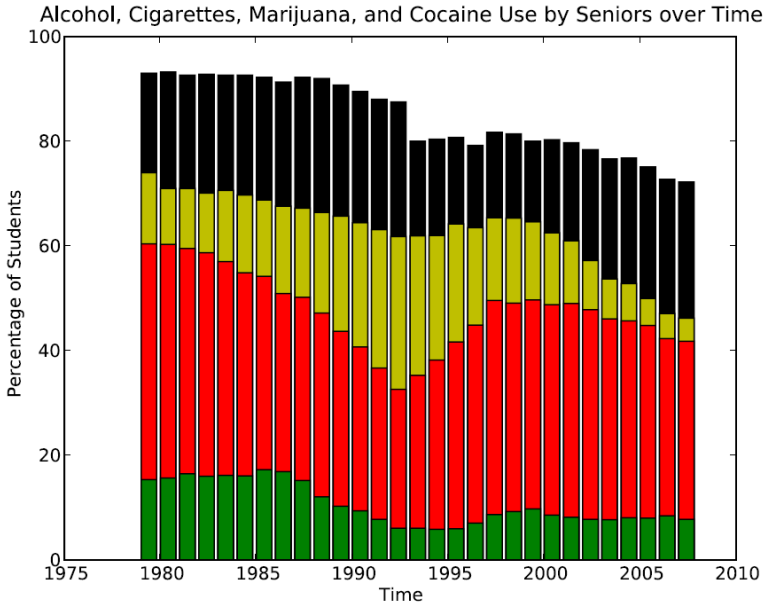
EXERCISES

1. Read Chapters 2 and 3 from the Tufte book. Pick a figure in Chapter 2 that Tufte argues is distorted but for which a correction is not provided. Why is this figure distorted? How should the distortion be corrected? In Chapter 3, Tufte discusses why artists draw graphics that lie. Considering that most scientists are not artists, in addition to the reasons listed in the Chapter, why else would a scientific "artist" produce a graphic that lies? Are any of these reasons justified (ethically)? (10 pts)

   Answers will vary. A good answer addresses all the questions and provides a reason other than those given by Tufte for explaining why/how a scientist might lie (e.g., save reputation, present only part of the data, filter data, etc.)

2. Define and save as a module a function that generates a series of bar charts on a single figure that plots the following drugs versus Year: Alcohol, Cigarettes, Marijuana, and Cocaine (the order is important to be able to visualize all the data). You can set a color for each bar graph using the keyword argument `color` (e.g., `color = 'k'`). The function should also label the axes and title the graph. Run you function in IDLE and save the result. What does this visualization tell you about the relationship among the use of these drugs over time? Is this a good visualization? Why or why not? (15 pts)

   Code answer is in file Exercise2_2.py



This visualization indicates the relative use of each drug through time. In doing so, it gives insight the relative quantity of use and allows you to compare the different trends (i.e., compare and contrast when peaks and valleys occur).

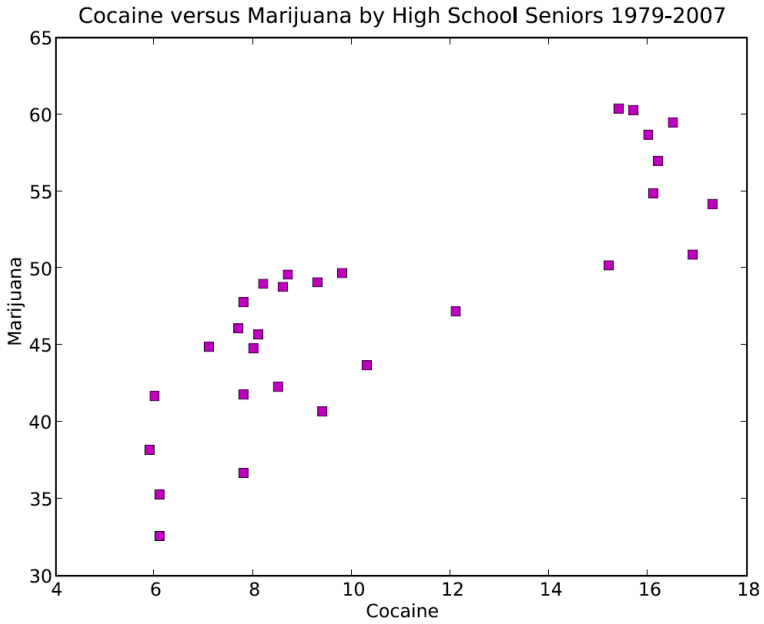
This is actually ***not*** a good visualization because it is a ***misrepresentation*** of the data. It has the appearance of being a "stacked" bar graph which implies that the amount of each type is proportional to the visible amount of bar. However, we know that we have actually just overlain a series of bar graphs and the value for Alcohol, for example should be measured from the axis to the top of the black bar, not just from the yellow to the black bar.

3. Define and save as a module a function that plots one type of drug data versus another (e.g., Marijuana versus Cocaine use). The function should have parameters that allow you to specify the two drugs to be compared and the format of the line (e.g., solid red line `'r'`, dashed blue line `'--b'`, magenta squares `'ms'`). The function should also label the axes and title the graph. Use your function to generate a variety of comparisons; save your two favorites and explain the relationship illustrated by each visualization. (15 pts)

   Code answer is in file Exercise2_3.py.

   A successful function includes parameters for identifying the data to be plotted on the x and y axis and a parameter for formatting the line. It should also take two parameters that provide strings so that the axes can be labeled appropriately. The goal of this exercise was to use parameters.

   Answers will vary depending on your visualizations. I hope, however, that some will not that this sort of comparison is best made with a scatter plot and not a line plot. Explanations should show some understanding that the graph is indicating a relationship between the two types of data and shows how changes in one correlate to changes in another. An example plot is below:



4. Define and save as a module a function that calculates the average drug use from 1979 to 2007 for each type of drug (except AnyDrug). You can return the calculated

values by creating a type of variable called an array that stores all of the results with a single name using the following syntax:

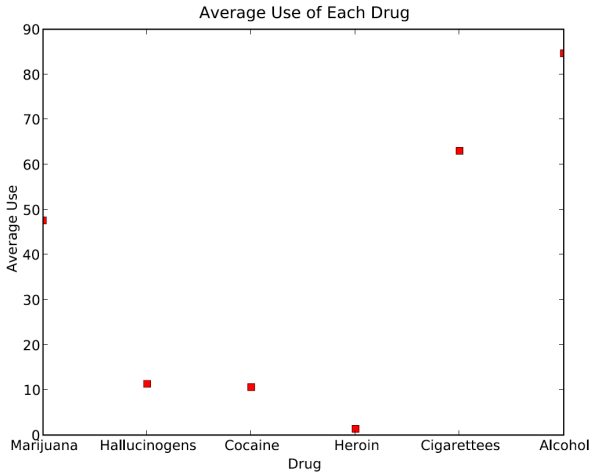
```
averageUse = [aMarijuana, aHallucinogens, aCocaine, etc...]
```

and then returning the `averageUse` variable. Test your function in IDLE. (25 pts)

Code answer is in file Exercise2_4.py

5. Define and save as a module a function that calculates the average drug use for each type of drug (except AnyDrug) and then plots the average drug use for each type of drug as a scatter plot using the `plot` command (which can generate univariate as well as bivariate graphs). HINT: you already did part of this!!! The function should also label the axes and title the graph. Labeling the ticks on the x-axis correctly is actually quite a difficult task and we will leave it for another day. When you print out your final plot, you can label the ticks manually with a pencil so that it is easier to interpret the plot. Or for the computer savvy you could edit the tick labels in a graphics editor or paint program. Save and run the module; save the resulting plot and explain the illustrated relationship. What insight does this visualization give you into the relative use of each drug by high-school seniors over the past 28 years? Is this a good visualization? Why or why not? (20 pts)

Code answer is in file Exercise2_5.py



Note that I have added the tick labels in this plot so that I can discuss it. Your plot would have labeled the ticks 1-6. This visualization strongly indicates the overall preference high school seniors have for alcohol, cigarettes and marijuana over all other drugs. It also emphasizes that in general cocaine use and hallucinogen use are very similar and that very few students use heroin. I wonder why?

No this is not a good visualization. While the plot does convey the information (i.e., the average values) it needs to, it does not do so effectively because it does not emphasize the main impact of the data, which is the relative magnitude of the

quantity that is average use. A bar graph would be a much more effective visualization. Also, it looks awful!