1. Real-time Rome question.

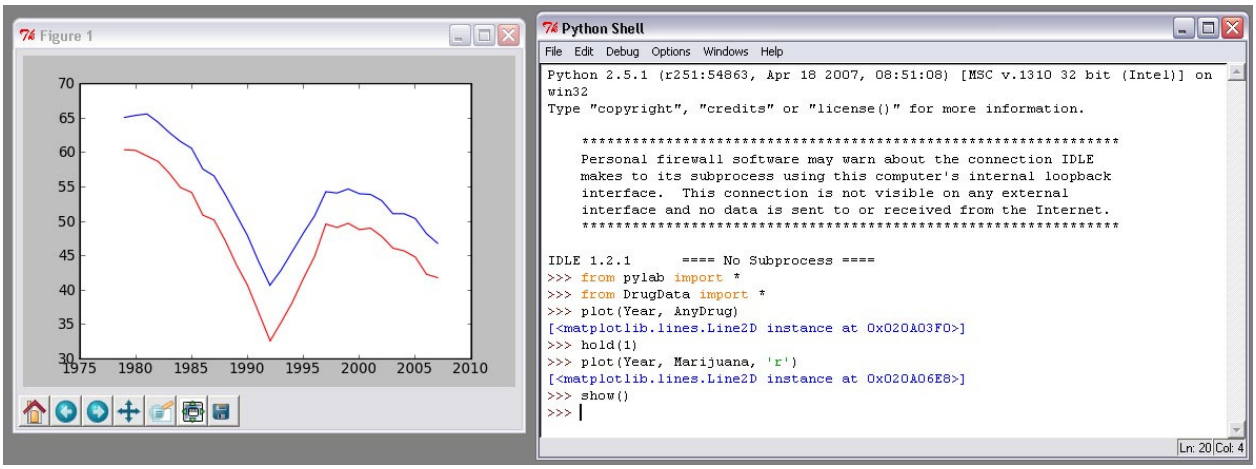
   Answers will vary. A good answer will address the whole question.

2. Plot the use of Marijuana versus time using a simple line plot. It would be interesting to compare the use of Marijuana through time to that reported for AnyDrug. To illustrate multiple plots on a single figure, we can use the command hold().

   Executing:

   ```
   >>> hold(1)
   ```

   tells Pylab to plot the next figure on the same set of axes that is already illustrated. To release the figure just execute the command `hold(0)`. Hold the figure and then plot AnyDrug versus time using a different color than the line for the Marijuana data. What does this comparison tell you about the use of Marijuana? Save the figure.



   This comparison tells you that the trend in all drug use over time exactly tracks the use of Marijuana over time and that the two are strongly *correlated*. (BTW: AnyDrug is all drugs except Alcohol and Cigarettes). The figure also implies that the most used drug is Marijuana. Thus, the figure implies that the pattern of overall drug use through time is mostly explained by the Marijuana data.

3. Repeat #2 using one of the other datasets. What does this comparison tell you about the use of that narcotic through time? Save the figure.

   Answers will vary depending on the comparison; but for all you should at least notice the following:

1. If you compared AnyDrug to Hallucinogens, Cocaine, or Heroin: ***that this visualization does not allow you to effectively make a comparison between overall drug use and the use of any of these drugs because of the y-axis scale*** (the percentage of folks using Hallucinogens, Cocaine, or Heroin is so little compared to the overall that the pattern gets smushed). For example, the figure implies that Hallucinogen use is fairly stable from 1985 – 1990 when in fact it shows the same cycle the overall drug use does! Similarly, Heroin use looks like it barely changes when in fact there is a very significant peak between 1995 and 2005!

2. If you compared AnyDrug to Cigarette use: up until the last few years, more high-school seniors smoked cigarettes than used any type of narcotic. Despite this the pattern of cigarette and drug use are remarkably similar (same valleys and peaks) making one wonder what is going on that is causing this pattern. Also, the pattern of cigarette use makes one wonder what happened between 2000 and 2005 to so markedly affect cigarette use but not to strongly affect narcotic use.

3. If you compared AnyDrug to Alcohol use: more students drink alcohol then use any type of narcotic. More striking, however, is that while the trend in the use of both is very similar (same decrease, peaks in valleys in same places); the spacing between the two never changes, implying that ***relative use of drugs and alcohol has remained consistent***!

4. Fancy it up a bit. Choose one of the narcotics and plot versus time. Modify at least three line properties to spruce up the image a bit. Save the figure. How does this manipulation affect the impact of and information conveyed by the visualization?

   Answers will vary. A good answer will 1) accomplish the task and 2) address how manipulating the line properties increased or decreased the information content and readability.

5. Pylab provides many other types of relational visualizations in addition to the simplescatter and line graphs generated by the `plot()` command. These include:

```
>>> bar(x,y)
>>> barh(x,y)
>>> stem(x,y)
```

   as well as means for visualizing a single variable:

```
>>> hist(y)
>>> boxplot(y)
```

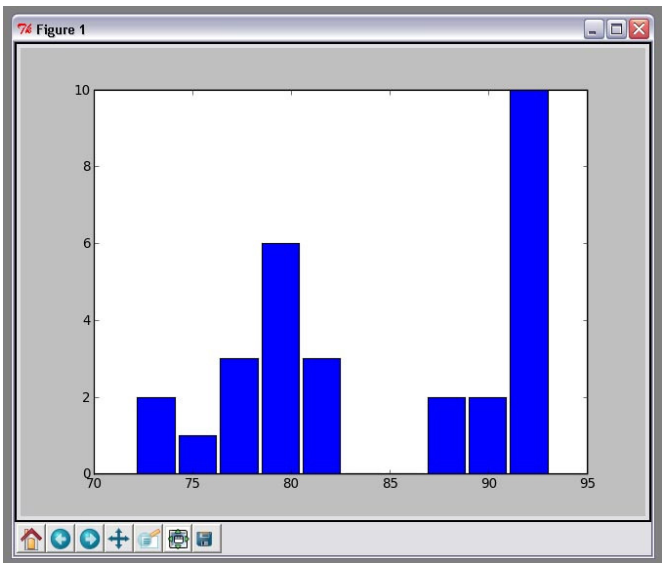
   Choose two of the drug variables and experiment with these types of plots. What type of graph does each produce? For the relational visualizations compare the drug usage data to time (Year) and for the single variable visualizations just use the drug usage data. How does the choice of visualization affect the message conveyed? Does it ever alter the type of the visualization (i.e., from Discovery to Communication or Insight)? Save your favorite graphics.

CMSC 120: Exercise 1: Answers

Answers will vary. The plot command generated either a line (which tracks a trajectory, in this case, through time) or a scatter (which emphasizes individual instances in time) plot. Thus, these two methods of plotting visualize the same data in very different ways. The goal of this question was to get you to think about how other types of plots changed the way the data were visualized and thus the information presented. I also wanted to see you consider whether or not a visualization was appropriate. I know that some types of plots (e.g., box plot) were unfamiliar.

Answers should include some aspect of the following:

1. **Bar plot**: plots data as a series of bars, where each bar is proportional to the value of the data. It represents data discretely; not really a good visualization of this data the bars detract from the point of the visualization (the trend through time) (more ink than necessary to convey the message)
2. **Barh plot**: sames as bar, but "x" data is on y-axis so that the bars are horizontal. Not really a good visualization for looking at a progress over time. Typically time is best viewed on the x-axis (there are some exceptions).
3. **Stem plot**: a scatter plot where the points are connected to the axis by a "stem" or line. Adds emphasis to the pattern without detracting from the message. I think this is a good visualization.
4. Histogram (**hist**): a histogram is a special univariate plot that displays the number of times each value was observed (in this case percentages of drug use through time). In the figure below, for example, I have generated a histogram of from the Alcohol data. The x-axis is the range of observed percentages per year, and the y-axis is a count of the number of times each observation was made. In other words, for example, there were ~6 years in which ~80% of the high school seniors used alcohol; 10 years when ~92.5% of the students used alcohol. Thus, the figure characterizes the overall expected amount of alcohol use. For our data, this is not a very good visualization as it really does not provide any insight into the system being examined; i.e., no useful information is conveyed.



5. **Box** plot: The box plot is actually a really neat visualization. It is a basically a diagrammatic version of a histogram. I am not going to explain the whole thing now (we will be looking at them in lecture), but for an example below I have again plotted the Alcohol data. The y-axis is the same as the x-axis in the histogram (the range of observed values). The box indicates where 50% of the data (observations) fall along the range. Notice how this makes it a continuous representation (instead of the discrete one in the histogram). The red line in the middle of the box is the median value. The vertical dashed line is just a graphical element added for continuity and the short horizontal lines indicate the boundaries within which (usually) 90% of the data fall. Again, not really useful for our data.