

36-315: Statistical Graphics and Visualization

Lab 11

Date: March 26, 2002

Due: start of class April 8, 2002

In this lab you will start the first phase of the final project, which is to formulate a statistical question about your state and try to answer it using the available census data. To do this, you will probably end up using most, if not all, of the techniques presented in class so far, including: histograms, dot plots, scatterplots, contours, surfaces, mosaics, and maps. If you look at the 200+ census variables for your state, you will find that they are arranged into eight basic groups:

1. Location
2. Population density
3. Ethnic composition
4. Age
5. Households/Families
6. Income
7. Education/Employment
8. Housing

For this lab, your task is to first pick a variable that interests you, e.g. PCTVACNT. Call this the ‘response’. Then try to explain the behavior of this response using variables from *three* of the other seven groups. For example, is it related only to income, or does ethnic composition play a role? Is ethnic composition irrelevant once you consider income? Do ethnic composition and income interact in predicting the response? Does the response correlate with geographical features of your state, like military bases and national parks? All of these questions would be pertinent.

Keep in mind the following points:

1. The data is at the census tract level, not the individual level.
2. Percentages are usually more informative than raw counts.
3. Know what the variables are really measuring. For example, PCTELEM means elementary education *only*, and ‘household income’ increases with the number of working people in the household.
4. What looks like a simple trend may actually be several distinct clusters.
5. Outliers can be just as interesting as the main trend, and should be investigated.
6. Many associations that appear significant can be explained by a lurking variable.
7. Let the data guide you to a conclusion, not the other way around.

You should justify your findings with plots, though the main emphasis here is learning about your state, not making plots. Your ‘grade’ for this lab will be an evaluation of how much progress you made toward understanding your response variable, and toward the final project as a whole. Your findings here do not have to be the ones you will use in the final project. It will probably be an iterative process.

We suggest that you *do use the scheduled lab time*, in order to clear up any problems you may have had in making plots in past homeworks, since these are exactly the kinds of plots you will be making again. If you don’t know how to make a particular plot, you are depriving yourself of what you could learn from that plot. As always, we do not promise to be available to give help outside of the lab hour.