The project as a whole is due at the end of the semester, but you will do pieces of it in weekly assignments, the first of which is due next Wednesday and is described on the last page of this handout.

The project is to describe how a variable of interest (the ‘response’) varies with respect to other census variables in your chosen state. Ideally, you would like to be able to predict the response in a census tract given any subset of the other census variables. For example, you would like to be able to predict rent from population density, rent from the percentage of homeowners, or from both. This requires knowing which variables are relevant to the response, how they interact, and special cases. Your project should provide useful information toward this goal, describing how the response varies with several of the census variables, with special attention to results that are non-obvious or surprising. The results should be understandable to a layperson, so you should not merely report the coefficients of a multiple regression.

We are hoping to have files ready with 2000 decennial census data for the project but we’ll begin with 1990 census data and switch when the 2000 data are available.

Pick a response from one of the eight variable groups listed below, and explain its behavior using variables from at least four of the other seven groups. You want to pick variables and analyses that show something non-obvious. An increase in income with education would be obvious, but an increase in income with family type (for a fixed education level) would be non-obvious. Outliers and changes in predictor strength (as seen on contour plots) are often surprising.

The report should follow a logical sequence and be reasonably concise (less than twenty pages). Throughout the report, you will probably have to make assumptions that cannot be verified from the available data. Make it clear which of your statements are assumptions and which come from the data.

You will probably find it useful to define your own tract categories, such as “desirable neighborhood”, based on the values of several variables. Then your results might simplify into rules like “if the neighborhood is desirable and vacancies are low, then rent is high, barring the following exceptions ...”
Grading

The report will be graded on the following criteria:

**Technique** Are your graphs well-designed? This includes, for example, variable transformation, smoothing parameters, symbols, aspect ratio, and color palette.

**Choice** Are your visualizations the right ones to support your argument? Does each tell a different story, or are they redundant?

**Interpretation** Is each graph interpreted completely and correctly?

**Depth** The variety and amount of information conveyed by your graphs (excluding redundant graphs). You want to go beyond simple trends found in scatterplots.

Other things to keep in mind:

1. The data are at the census tract level, not the individual level. Thus the classroom discussion about causal conclusions is relevant. As you develop causal interpretations be sure to also draft a paragraph or so explaining why one needs to be cautious about their use.

2. Percentages are usually more informative than raw counts.

3. Know what the variables are really measuring. For example, PCTELEM in the 1990 files means elementary education *only*, and ‘household income’ increases with the number of working people in the household.

4. What looks like a simple trend in a graph may actually be the overlay of several distinct clusters.

5. Outliers can be just as interesting as the main trend, and should be investigated.

6. Many associations that appear to be “significant” can be explained by a lurking variable.

7. Let the data analysis guide you to a conclusion, not the other way around.
First Project Assignment  
Due: April 19, 2006

The first step of the project is to choose the dataset that you want to use, the response variable that you want to study, a set of relevant predictor variables, and questions to answer. At http://www.stat.cmu.edu/~fienberg/Stat36-315/project/project.html you will find links to short and long descriptions of the census variables available. The 200+ census variables 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

Your task is to first pick a variable that interests you, e.g. PCTVACNT. Call this the ‘response’. Then identify useful predictor variables from four of the other seven groups. Give a short list of the questions you could answer using these variables. You are welcome to make plots to help in your search, but no plots need to be handed in.

For example, some questions you might try to answer in the project are: is the response 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?

There are various places where you can find inspiration. Throughout the class, various outliers and unusual trends were explored in Pennsylvania. You can choose to pursue one of these in more depth for your state. For example, as we saw in class, Kentucky had an East-West dichotomy on several variables, and that this was partly explainable by geography. For some states, the distinction is a major city versus the rest. Does your state have similar dichotomies, and are they also explainable by geography?

Examples of questions that people express interest in:

- Are educated people more likely to be married?
- What demographic groups are most likely to be married?
- Do Hispanics have larger families?
- Do larger households make more money? What if many of them work?