36-315: Statistical Graphics and Visualization

Final Project
Date: April 16, 2002 **Due: May 6, 2002**

The final project is to describe how a variable of interest (the 'response') varies with respect to other census variables in your state. Perhaps the highest level of understanding one can have about a response variable is if one could make a decent prediction of the response given any subset of the other census variables for a census tract. For example, predicting rent from population density, the percentage of homeowners, or 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 multivariate regression.

Pick a response from one of the eight variable groups listed in lab 11, and explain its behavior using variables from at least four of the other seven groups. You do not have to use the same variables that you used for labs 11 and 12. 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. Interactions and outliers are often surprising.

The report should begin with a 2-3 page "Executive Summary" which explains in layman's terms what you have found, and why it is interesting. There should be no graphs in this summary. The report will then contain a longer technical description of what analyses you performed and how you arrived at your conclusions. The report should follow a logical sequence and be reasonably concise. Try not to exceed 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 ..."

The report will be graded on the following criteria:

Visualizations How effective are your visualizations. Are they the right choice of visualization for supporting your argument.

Logic How well your conclusions are supported by the data. How well your visualizations fit together to make an argument.

Novelty How non-obvious are your findings. That is, how big of a role did the data play in obtaining them.