Boston Property Values:

Why Some Areas Cost More

By Dan Schult

This document is meant as an example of a Methods and Results Section to which students can write an accompanying Discussion/Conclusion Section.

(Brief) Statement of the Problem

What makes property values go up and down? Which factors determine the property values of houses? Are there some concrete steps a town can take in order to raise property values by manipulating these factors? These questions are longstanding issues that local political leaders have had to confront over the years.

Property values for houses depend on many factors. Most of these factors can be thought of in terms of factors that make a location desirable. More desirable areas have presumably higher property values. But, other factors may be used as signals to potential buyers about what an area is like. For example, consider tax rates. While high tax rates, in isolation, make a location undesirable, they signal to buyers that wealthy people live in this location. The wealthy neighbors presumably attract more wealthy buyers, potentially increasing property values. Factors such as these don t directly affect the desirability of the location, but may indirectly affect property values because of what they say about the community.

This study is intended to explore which factors are most strongly related to property values. The results will presumably help us answer some of the questions raised above.

(Brief) Background

Many previous studies have looked at the issue of property values.
[A good background would then go on to describe what some of the other studies have done and perhaps say what was lacking from those studies that will be improved here.]

One of the difficulties with studying factors affecting property values is the variation from state to state. To study national data, we would probably need to control for which state the data is from. That could be an important factor affecting property values that towns could not change and therefore would not be important for this study. To alleviate this potential problem, we use data for census tracts within one area (the Boston area).

Methods

We use data from the 1970 census on property values and many other factors for each census tract in the Boston area. This data is available through the StatLib[1] collection of databases for educational purposes. Variables describe geographical information such as the distance to major interstates and employment centers. They also describe zoning information such as the proportion of the lots with large lot size and the proportion of land set aside for industry and the average number of rooms per dwelling. Economic variables are the poverty rate of the tract, the property tax rate, the pupil to teacher ratio in the school district (used as a measure of money spent on education) and, of course, the median property values of owner-occupied homes. The complete list of variables and descriptions are in table 1.

Variable	Description
Crime Rate	Per capita crime rate by town
Big Lots	Proportion of residential lots over 25,000 sq. ft.
Industry	Proportion of industry acres per town
Pollution	Nitric oxides concentration
Rooms per House	Average number of rooms per dwelling
Age of Tract	Proportion of owner-occupied units built prior to 1940
Distance to Work	Weighted distance to five major Boston employment centers
Access to Highways	Index to accessibility to radial highways
Tax Rate	Full-value property tax rate
Education Value	Pupil-teacher ratio by town
Minority Percent	1000(Bk-0.63)^2 where Bk is the proportion of blacks by town
Poverty Rate	Percent of population under poverty line
Property Value	Median value of owner occupied homes

Table 1

The data in this dataset had some obvious defects. Of the approximately 500 census tracts, none had median property values listed above $50,000 and 15 tracts were listed as exactly $50,000. We assume that the top value entered into the database was $50,000 and these 15 tracts actually had values larger than that. To remove any problems caused by this artificial cutoff, we exclude those 15 tracts of data and assume that the relationships between factors for excluded tracts are the same as those for the remaining data.

[1] http://lib.stat.cmu.edu/datasets /boston.htm on January 10, 2002.

The correlation coefficient between each variable and median property values was computed. By looking at the absolute value of these coefficients, we identified the two variables with the most correlation (positive or negative) to property values. Regression lines were then obtained for these two variables - poverty rate and average number of rooms per house.

The regression graphs allowed us to identify three potential outlier points in the data for rooms per house. The largest and two smallest values did not fit the rest of the data. It is quite possible that the data was typed incorrectly, or that these three tracts are special in some respect and should be treated differently. We did not remove them from our database, however, because we have no evidence that they are faulty data except that they don t seem to follow the trend of the other data.

The residual plot for poverty rate shows a potential nonlinear relationship between property values and poverty rate. To correct this, we can look for the best parabola through the data by using a multiple regression of property values on poverty rate and poverty rate squared. This was done in Excel by creating a new column in the dataset and using a formula to square the poverty rate values.

Using the regression tool in Excel, we were able to perform a multiple regression including the quadratic poverty variables and the variable Rooms per House . Multiple regression allows us to control for one variable while finding the correlation with the other variable. Thus the slope of resulting regression represents how much property values increase as the poverty rate increases while holding the rooms per house constant. This is a useful way to control for potential confounding. In this case, plotting a regression line is not reasonable because it would be a regression plane in a three dimensional picture. But residual plots against each variable are still useful for identifying outliers and other features of the dataset.

Results

The correlation coefficient between each variable and median property values appears in Table 2.

Variable Name	r
Crime Rate	-0.45
Proportion of housing lots over 25,000 sq.ft.	0.40
Proportion of industrial acres per town	-0.60
Boundary of Charles River? (Yes/No)	0.07
Pollution levels (nitric oxide concentration)	-0.52
Average number of rooms per dwelling	0.69
Distance to five major Boston employment centers	-0.49
Access to radial highways (index)	0.37
Tax rate	-0.48
Pupil-teacher ratio by town	-0.52
Proportion of minorities by town	0.36
Poverty Rate	-0.76

Table 2

The largest correlation in absolute value was for poverty rate (r=-0.76). The regression line of property values on poverty rates has slope –0.84 and intercept 32.54. A scatter plot appears in figure 1 with the residual plot in figure 2.

The second highest correlation was for the average number of rooms per dwelling (r=0.69). The regression line had slope 8.27 and intercept –30.01. The scatter plot and residual plots for this variable appear in figures 3 and 4. Notice that the highest and two lowest values for Rooms per House are potential outlier points.

The multiple regression for property values against poverty rate, poverty rate squared and rooms per house yields three slopes and an intercept. The intercept is 12.89 while the slope for rooms per house is 3.67, the slope for poverty rate is –1.59, and the slope for poverty rate squared is 0.0296. The residual plots appear in figures 5, 6 and 7.

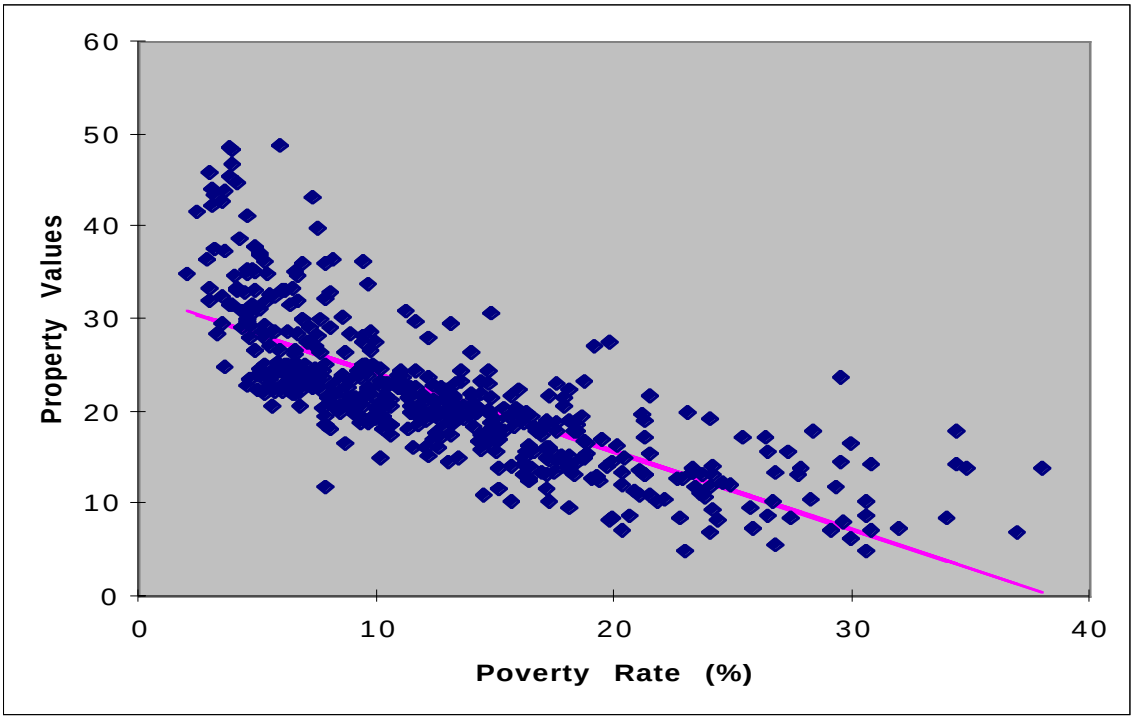

Figure 1

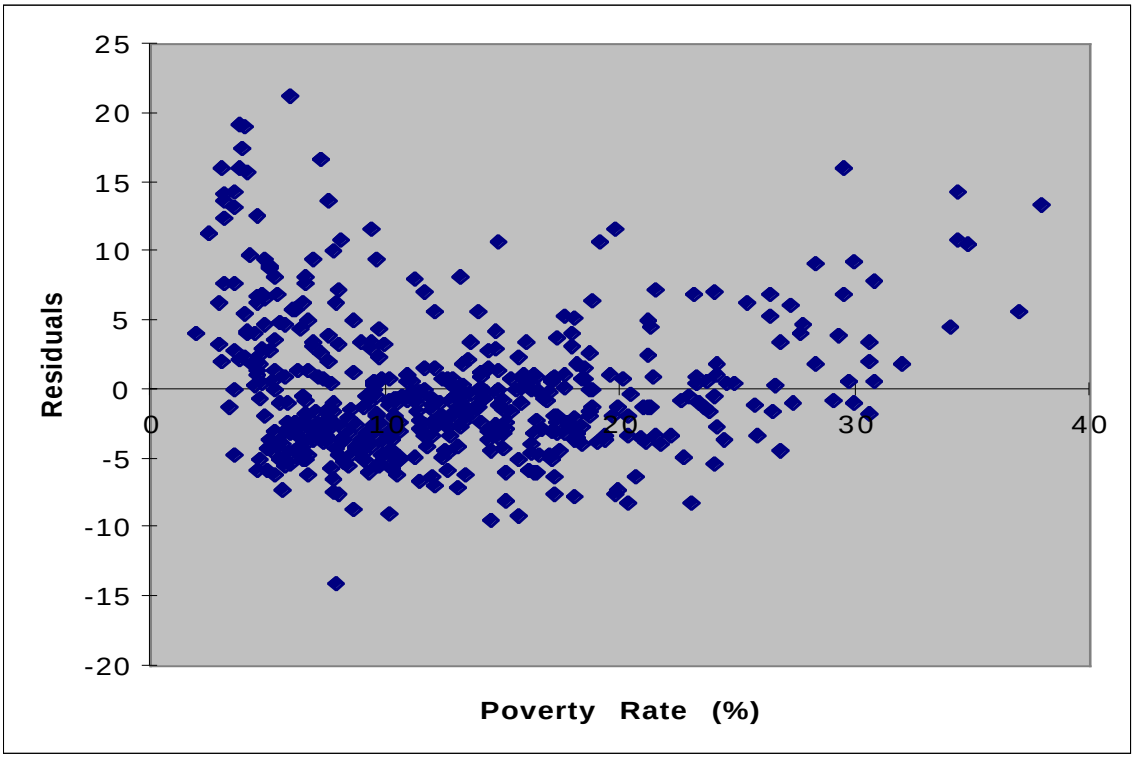

Figure 2

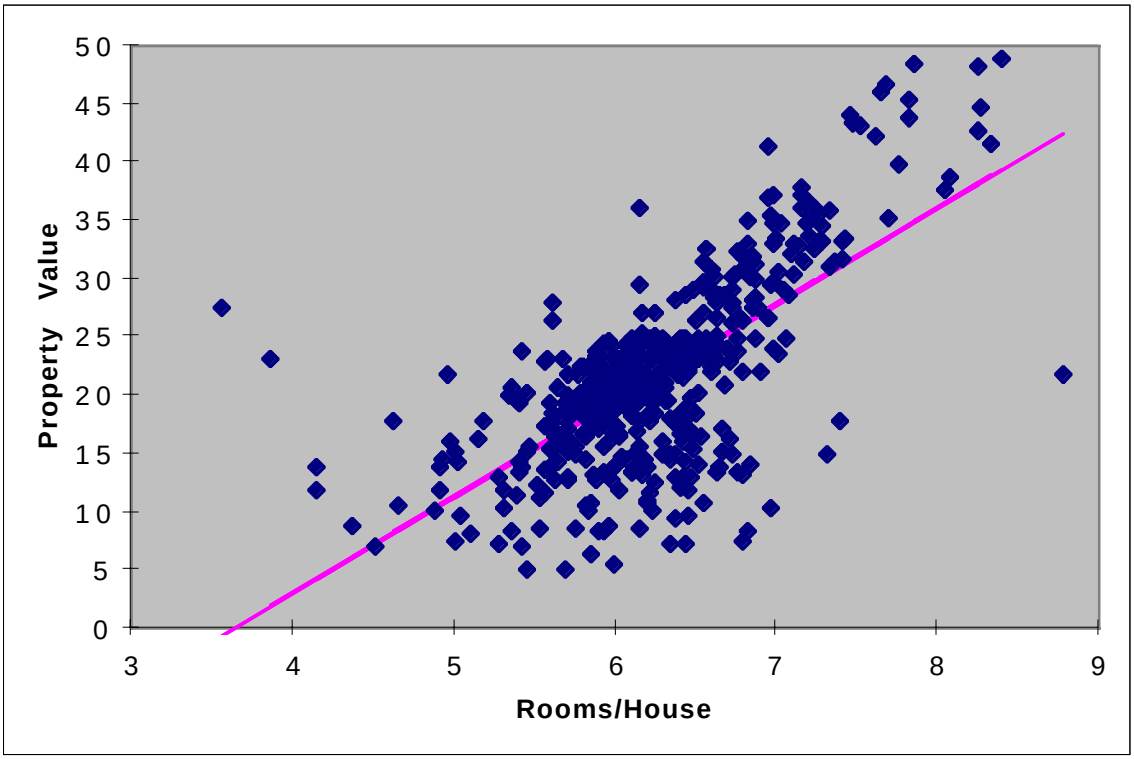

Figure 3

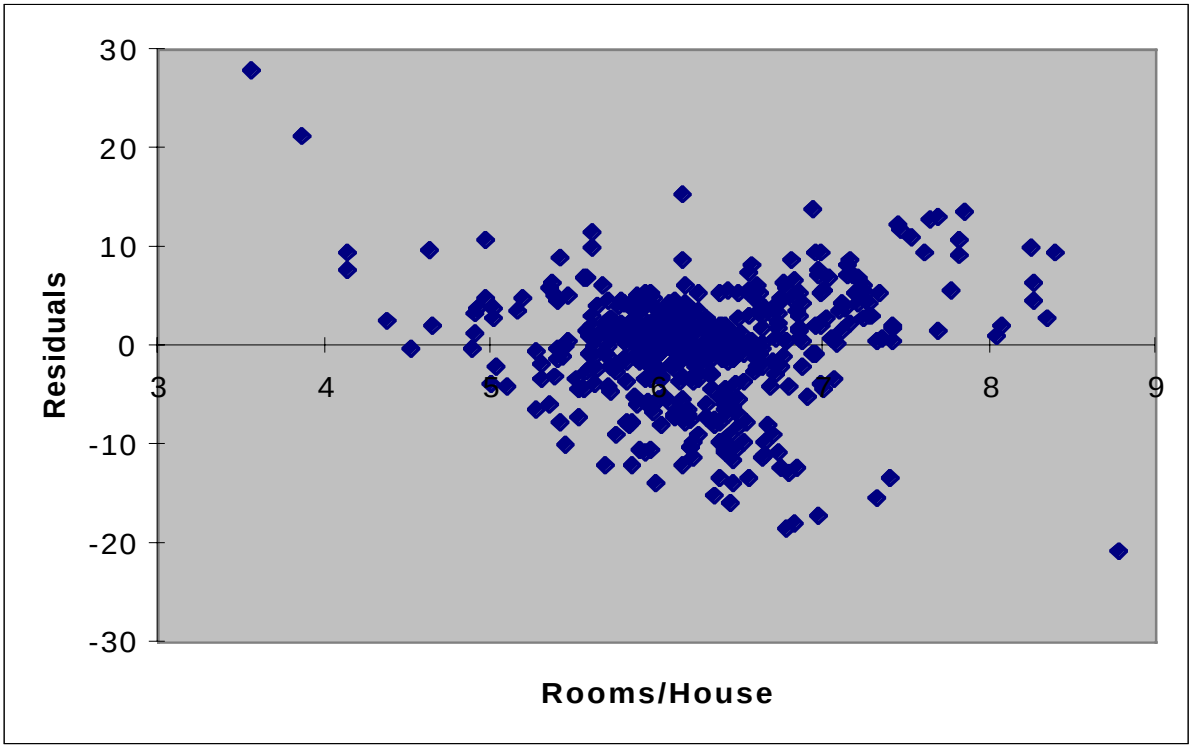

Figure 4

Figure 5

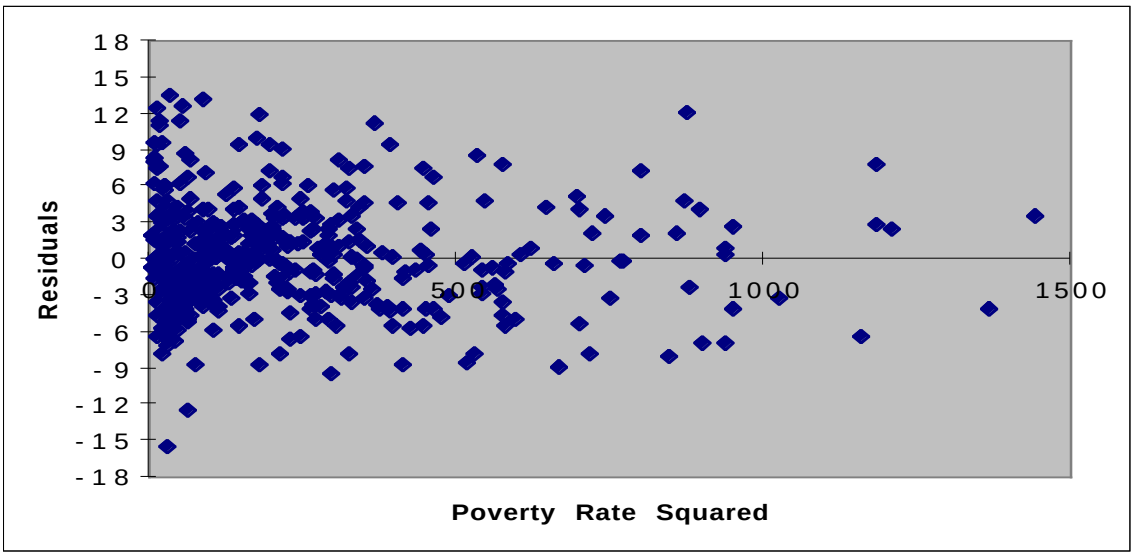

Figure 6

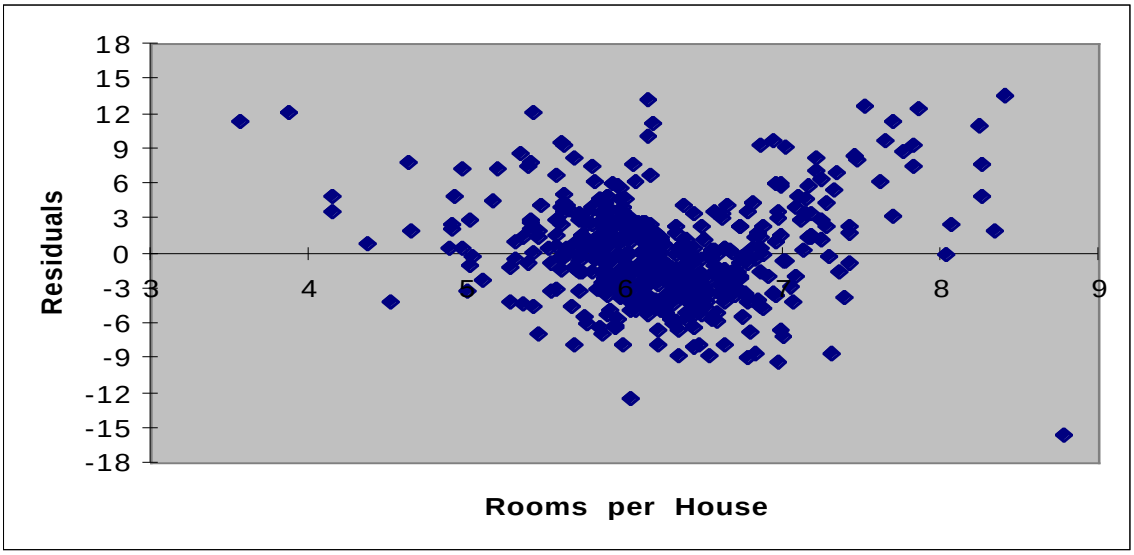

Figure 7