In-Class Project 3

Submit in group. Due March 27th.

The third in-class project aims to implement regression analysis in Excel to analyze data and make inferences. I have uploaded a data set in Canvas under the folder “Data”, which contains 291 housing sales in Lexington. You need to use this data set to analyze, run several regressions, and make statistical inferences about your results.

1 Data

The data set contains 291 sales. Each sale records detailed information of house attributes, including its sale date (saledt), sale price (price), building square footage (bldgsqft), year built (oldest), number of bedrooms (bedrooms), number of full bathrooms (fullbaths), number of half bathrooms (halfbaths), whether or not it is a full brick house (allbrick), whether or not it is a partial brick house (partbrick), and its sale year (saleyr).

In addition, it also provides distance to urban service boundary (dist2urban), distance to city center (centdist), distance to its public high school (dist_school), and matched high school ACT test scores.

2 Descriptive Statistics (15 points)

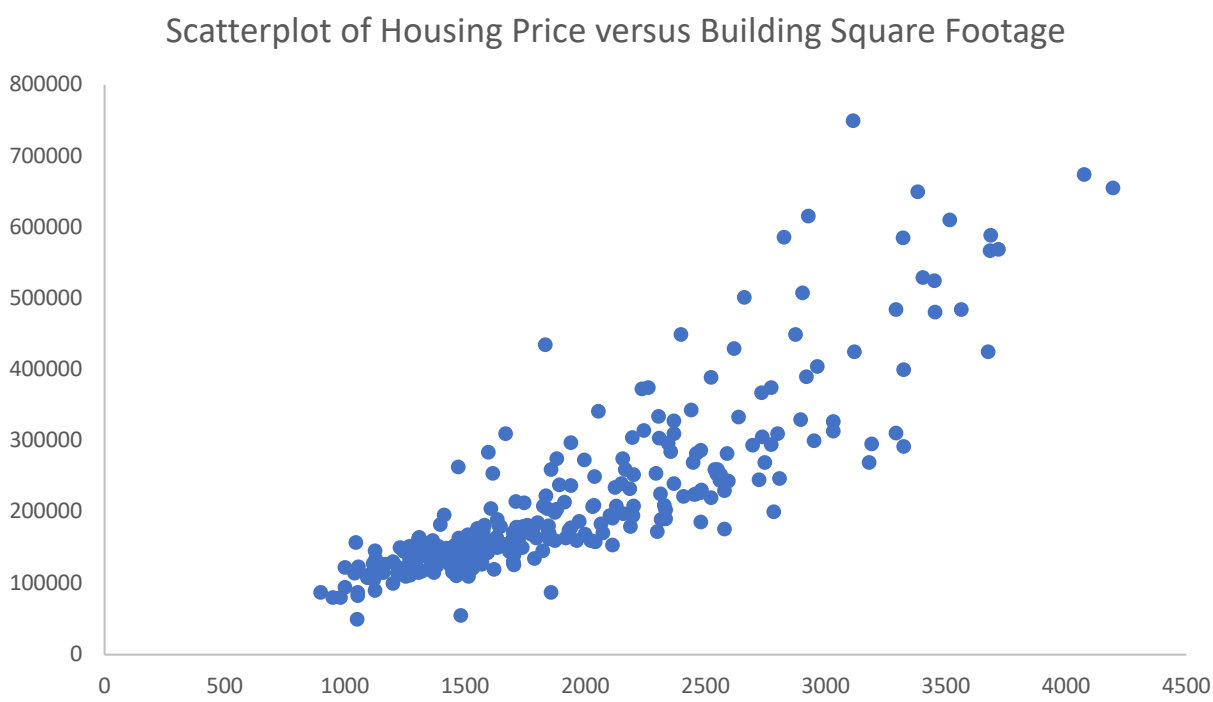
First, as always, we need to summarize the data and have a big picture about the major variables. That is to say, I want you to make a summary statistics table like what we did in previous projects. The table should include mean, standard deviation, minimum, and maximum for the following variables:

1) House characteristics: price, bldgsqft, age, bedrooms, fullbaths, halfbaths, allbrick.
2) Distance measures: dist2urban, centdist, dist_school.
3) School quality: English, math, reading, science, composite.

3 Graphical Evidence (15 points)

Then we plot some graph to see correlation between housing price and some variables we think could affect housing price. In this project, **I want you to make scatter plots for building square footage, bedrooms, and full bathrooms**. Then discuss what relationship you find in these graphs.

For example:



It looks like a positive linear correlation.

4 What Determines Housing Price?

1) We first examine the impact of some house attributes on housing price. Consider the following linear regression model:

$$price = \beta_0 + \beta_1 bedrooms + \epsilon$$
$$price = \beta_0 + \beta_1 fullbaths + \epsilon$$
$$price = \beta_0 + \beta_1 halfbaths + \epsilon$$
$$price = \beta_0 + \beta_1 bedrooms + \beta_2 fullbaths + \beta_3 halfbaths + \epsilon$$

Report the Excel output for each regression and write down the regression equation. Interpret the regression statistics and coefficient of variables. Discuss the statistical significance and economic significance for all variables.

Compare the four models in terms of goodness-of-fit, and coefficients of major variables. (40 points)

For instance, here is an example I did for the first model:

Regression Statistics	
Multiple R	0.63912708
R Square	0.40848343
Adjusted R Square	0.40643666
Standard Error	76687.3495
Observations	291

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-158589.34	23902.1843	-6.634931	1.5979E-10	-205633.78	-111544.91
bedrooms	96895.4757	6858.84223	14.1270892	8.1773E-35	83395.8583	110395.093

The regression equation then is

$$\widehat{price} = -158589.34 + 96895.4757 bedroom$$

The intercept is -15859.34, meaning if there is 0 bedroom for a house, no buyers would be willing to purchase it. Actually, the value for such a house is negative, so people need to be reimbursed to live in such house.

The coefficient for bedrooms is 96895 means **holding other things constant**, **increasing 1 bedroom** will lead to a 96895$ increase in housing price. And it is statistically significant (statistically different from zero), given t-Stat is large and greater than the critical value, or P-value is close to zero, or the 95% confidence interval does not include zero.

2) Now, **add building square footage** into the preferred model you choose in previous section. For example, if you prefer model 1, now you should run

$$price = \beta_0 + \beta_1 bedrooms + \beta_2 bldgsqft + \epsilon$$

Report the regression output from Excel, and discuss goodness-of-fit, coefficients, and interpret. (15 points)

What do you find in this model and how do you explain it? (15 points)