Real estate agents know the three most important factors in determining the price of a house are *location*, *location*, and *location*. But what other factors help determine the price at which a house should be listed? Number of bathrooms? Size of the yard?

A random sample is drawn from publicly available data on thousands of homes in upstate New York. The sample data are saved in the file '*real_estate.csv*'. The variables include the price (in dollars), total living area (in square feet), number of bathrooms, number of bedrooms, size of lot (in acres), age of house (in years), etc.

The first part of the project aims at examining how well the selling price can be predicted using the home size, as measured by living area. Analysis of nontrivial data entails lengthy calculations, which should be performed by utilizing statistical software. You are required to manage the data set and to obtain important results from data analysis through effective use of R. At the same time, you may appreciate the role of linear regression models in scientific inquiry.

In your report, make sure you

(1) Construct a scatterplot of the selling price (Y) against the living area (X). Comment on the association between the two variables.

(2) Explain why the selling price is selected to be the response variable.

(3) Regress the selling price on the living area. State the estimated regression function; and superimpose it on the scatterplot in part (1).

(4) Report the values of MSE and R^2. Interpret R^2 in the context of the problem.

(5) Obtain a 90% confidence interval for β_1, and interpret it in the context of the problem.

(6) Test whether β_1 is significantly different from zero at the 0.10 level of significance. State the alternatives, the value of t-test statistic, p-value, and your conclusion.

(7) Predict the selling price for a home with 1850 square feet. Obtain a corresponding 90% prediction interval.

(8) Superimpose the (pointwise) 90% prediction band on the scatterplot in part (1).

(9) Construct a scatterplot of semi-studentized residuals against X and a normal probability plot of semi-studentized residuals. Use the two plots to perform model diagnostics.

(10) Comment on the possible simple transformations of either Y or X, or of both, if any, that can make a linear model more appropriate for the transformed data. *Note: there is no need to refit the model to the transformed data.*

IMPORTANT NOTES

- You are required to work independently. Doing otherwise risks charges of cheating and hence a grade of F.
- Your answers should be typed up on letter-size papers, double spaced and with a font size of

at least 11 points.

- All numerical results should be obtained from R outputs/commands; no hand-calculations. No other statistical software but R should be used.
- **Include only the relevant computer output**, if any, to justify your answers.

GRADING CRITERIA

- Correctness of analysis results (and interpretations, if any), quality of graphs and tables, and clarity of the report.