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 second part of the project aims at building a linear model to predict the selling price using a set of potential predictor variables. 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 include the analyses for the following models.

Model #1*: a multiple regression model including the living area (X_1), number of bedrooms (X_2), and number of fireplaces (X_3) as predictor variables.*

(1) Obtain the scatter plot matrix and the correlation matrix of all the four variables (that is, the response and the three predictors). Is there significant evidence of multicollinearity?

(2) State the LS regression function and interpret the coefficient of the number of bedrooms (X_2).

(3) Conduct an F test of overall linear relationship. State the alternatives, the value of F-test statistic, p-value, and your conclusion.

(4) Compute the extra sum of squares $SSR(X_3|X_1, X_2)$ and the coefficient of partial determination $R^2_{Y3|12}$.

(5) Conduct an F test of whether X_3 is helpful, given that X_1 and X_2 are in a model. State the alternatives, the value of F-test statistic, p-value, and your conclusion.

Model #2*: a multiple regression model including the living area, "Central Air" (an indicator variable coded as 1 if a house has central air conditioning, 0 otherwise), and their interaction term as predictor variables.*

(6) Construct a scatterplot of the selling price against the living area for the two groups: *Central Air* = 1 and = 0. Explain why the interaction term is included.

(7) State the LS regression function and interpret the coefficients of *Central Air* and the interaction term, respectively.

(8) Conduct a partial F test for identical regression function. State the alternatives, the value of F-test statistic, p-value, and your conclusion.

***Model #3**: building the "best" regression model using all potential predictor variables (with no interaction terms).*

(9) Obtain the best model using the "backward elimination" procedure. Report the subset of predictor variables to be included in the model.

IMPORTANT NOTES

- You are required to work independently. Doing otherwise risks a charge of cheating, which may result in a grade of zero.
- No other statistical software but R should be used.
- All numerical results should be obtained from R outputs; no hand-calculations.
- The report should be typed up on letter-size papers, double spaced and with a font size of at least 11 points.
- There are no specific requirements about the format of the report; it is up to you.
- Include the relevant computer output in the report.
- Submit R code in a separate file.

GRADING CRITERIA

- Correctness of results of analysis and interpretations, quality of graphs and tables, and clarity of the report.