# Project Proposal

**Substantive Question:** Option 1

How do the variables Year of construction (YRCON), Number of floors (NFLOOR), Number of underground floors (BASEMNT) and Floor to ceiling height (FLCEILHT) influence Whether there are elevators or not (ELEVTR) in the building according to the 2012 US Commercial Building Energy Consumption Survey (CBECS) data? If they are related, predict “Elevators or Not” based on the analysis.

**Specific Data Sets and Variables:**

The data should come from the 2012 US Commercial Building Energy Consumption Survey (CBECS). The dependent variable is “whether there are elevators or not” named ELEVTR in the data set. The independent variables are Year of construction (YRCON), Number of floors (NFLOOR), Number of underground floors (BASEMNT) and Floor to ceiling height (FLCEILHT).

**Analysis Methodology:**

Since “Whether there are elevators” is a typical binomial distribution variable, it is appropriate to use logistic regression to analyze its factors. The plan of analysis is shown below.

(1) Firstly, data needs to be filtered and cleaned and only valid information is retained. In order to facilitate subsequent prediction and testing, the overall data needs to be divided into training set and test set according to a certain proportion.
(2) Secondly, perform logistic regression and judge whether each independent variable has sufficient influence on the dependent variable based on the p-value. Also, eliminate insignificant factors to improve the prediction accuracy.
(3) Thirdly, predict “whether there are elevators based on the logistic regression results and test the accuracy of the prediction.
(4) Fourthly, in order to make the logistic regression more reasonable and the prediction more reliable, cross-validation should be used to compare the models. The data will be divided into 10 folds.
(5) Fifthly, in order to show the relationship between the variables, choose the best performing one out of 10 folds in the cross-validation, that is, the one with the highest accuracy.
(6) Lastly, draw ROC diagram and calculate AUC value to test its performance.

**Statistical Software:**

R codes will be applied in this project to complete all the analysis.