

IE Projects: Step 2.

In all the projects the ultimate goal is to build a linear regression model with some or all of the predictors available to you. In all cases you'll eventually do some model building as you may not need to use all variables (or it may not be wise to use all variables). It might also be necessary to transform some of the X 's and possibly transform the Y .

The next step in the project is to explore the data in some detail in advance of fitting regression models. Primarily we want to get a sense of the data, see if there are any extreme values/outliers we need to worry about and also to consider

- Might we need to transform a Y or an X ?
- Do we want to eliminate some X 's which are highly correlated with other X 's

More specifically, what you should do is

1. Get descriptive statistics on each of the variables (each X and the outcome).
2. Get correlations among all of the pairs of variables. You'll want to get both Pearson and Spearman correlations.
3. For each of the predictor values that is quantitative, plot Y versus X .
4. For each of the predictor variables that are categorical/ qualitative, plot Y versus X and get summary statistics for Y over each of the possible categories of X .