Case Study: Predicting House Sale Prices

The goal of this study is to construct a model for predicting the selling price of a house based on house properties, such as size of the house (in square feet), number of bedrooms, and other characteristics. The houses in this study are in the Seattle area, with sale dates ranging from May 2014 to May 2015. This data is contained in the file "Seattle.csv". The total number of records in the dataset is 613 houses.

Following is a list of the variables, along with variable descriptions:

Variables	Description	Data Type
id	Identification number for house	Numeric
date	Date house was sold	String
price	Sale price (in US dollars)	Numeric
bedrooms	Number of bedrooms	Numeric
bathrooms	Number of bathrooms	Numeric
sqft	living square footage of the home	Numeric
sqftlot	square footage of the lot	Numeric
floors	Total floors (levels) in house	Numeric
waterfront	House which has a view to a waterfront	Numeric
view	Has been viewed	Numeric
condition	Overall condition of property: 1 = worn out property; 5 = excellent	Numeric
grade	Overall grade given to the housing unit, based on county grading system. 1 = poor,13 = excellent	Numeric
sqftabove	square footage of house apart from basement	Numeric
sqftbasement	square footage of the basement	Numeric
yrbuilt	Year house was built	Numeric
yrrenovated	Year house was renovated	Numeric

Some notes:

- You may want to do a little research (not too much) on how the predictors above would be expected to affect selling price.
- This data may require some cleaning and "data snooping". Before you do anything, you need to look at plots and summaries for any potential problems. For example, if 99% of the data is "Category" A, and only 1% "Category B" for a predictor, perhaps including that predictor would be ill-advised.
- You may also want to change the variable type for some predictors. Variables that look like "integer" objects to R may need to be changed to "factor". Also, some numeric variables could be considered as factors.