Housing EDA

David Gerard

2019-09-19

Researchers were interested in predicting residential home sales prices in a Midwestern city as a function of various characteristics of the home and surrounding property. Data on 522 transactions were obtained for home sales during the year 2002. The 13 variables are

- `Price`: Sales price of residence (in dollars)
- `Area`: Finished area of residence (in square feet)
- `Bed`: Total number of bedrooms in residence
- `Bath`: Total number of bathrooms in residence
- `AC`: 1 = presence of air conditioning, 0 = absence of air conditioning
- `Garage`: Number of cars that a garage will hold
- `Pool`: 1 = presence of a pool, 0 = absence of a pool
- `Year`: Year property was originally constructed
- `Quality`: Index for quality of construction. `High`, `Medium`, or `Low`.
- `Style`: Categorical variable indicating architectural style
- `Lot`: Lot size (in square feet)
- `Highway`: 1 = highway adjacent, 0 = highway not adjacent.

The data are available in "estate.csv" at https://dcgerard.github.io/stat_412_612/data/estate.csv.

Perform an exploratory data analysis to come up with some hypotheses. Some suggested ways to focus your research:

- Which variables are categorical? Which are quantitative?
- Change the values of the categorical variables to something more informative.
- What variables are marginally associated with price? Use plots and summary statistics.
- What variables are marginally associated with each other? Use plots and summary statistics.
- If a variable is marginally associated with price, are there some other variables that could explain that association? Use plots and summary statistics.
- Does there appear to be any discrete groupings of houses?
- Are there any unusual observations?
- What transformations should you perform to make associations more linear?
- Try making new variables based on existing variables.
- What variables should be discretized (or have values aggregated) because there are too few values and/or the association seems discrete?
- If you know linear regression, try out the `lm()` and `step()` functions to choose a tentative model.