Data Story 5: Quantifying Factors the Control Housing Sales Princes

George I. Hagstrom

Assignment

You have been given a dataset consisting of a single file, `ames_prices.csv`, and a data dictionary `data-dictionary.txt`. This dataset contains the sales price of each home as well as other attributes of the home and the sales transaction, such as the lot size, the square footage of each floor, when the home was built and remodeled, whether the kitchen is upgraded, etc. A data dictionary has been included with the data.

These variables are closely associated with the sales price of the home. This dataset is a model dataset for demonstrating machine learning techniques and for prediction.

Here our goal is to use a simple statistical model to tell a story about the factors that affect housing prices. I want you to build a linear regression/generalized linear model that incorporates several variables that influence housing prices (you should include the most influential factors on housing prices such as square footage, overall quality, lot size, type of sale, etc, but also pick a few smaller factors whose marginal effects can be calculated and plotted such as garage size or kitchen quality), and then make model visualizations to tell your story. I highly encourage you to use the `marginaleffects` package to help with making visualizations. If you see fit, you can also use dimensionality reduction techniques to improve your understanding of the dataset.

Tips and Information About the AMES dataset

This dataset was downloaded from kaggle, which holds periodic introductory competitions using this data. It was originally compiled by De Cook for educational purposes. Skim this article for some tips on using this dataset: De Cook 2011. De Cook makes some suggestions for simplifying the data that you can take if you need to. The assignment is much more about using visualizations to explain your model and its meaning than it is about making a fancy model or predictions, if you are feeling stuck with modeling or feeling analysis paralysis feel free to see how others have modeled this dataset. I recommend using simple models (linear

regression and close variants), but if you feel like you can make visualizations showing the importance of individual factors using other modeling frameworks you may use those other models.