Summary of EDA of the Ames Housing Dataset

Background and Motivation

This analysis was performed on the Ames, Iowa housing dataset, made available via the "House Prices - Advanced Regression Techniques" Kaggle competition (Kaggle 2022). The dataset contains records of every home sale in that town from 2006-2010, along with 80 other explanatory variables. This data allows researchers to interrogate and understand which housing characteristics are most influential in predicting home sale price. The Kaggle competition already splits the data into training/test datasets: all discussion of data below is limited to the "training" dataset, which contains 1,460 observations.

This research has a strong financial motivation: the residential housing market is not efficient, so having a quantitatively defensible expectation of a home's true value can confer a meaningful advantage during purchase negotiations. Furthermore, an expectation of return on investment could guide homeowners' decisionmaking when considering home renovations.

Finally, this dataset spans the financial crisis of 2008-2009 which saw a material reduction in home prices across the nation (Indiviglio, 2010). This dataset provides a window into the impact the global financial crisis had on a small midwestern town's housing market.

Description of Data

Dependent Variable

The dependent variable of this dataset is "SalePrice", the amount of money changing hands when the property was sold. Sale prices range from $35k-$755k, and exhibit right skew: the median sale is $163k while the average sale price is $181k.

There is justification to log-transform the sale price. This effectively changes the relationship between two values from being additive (absolute dollar difference) to being multiplicative

(approximately percentage difference). This makes intuitive business / practical sense, and also makes the dependent variable much closer to a normal distribution.

Independent Variables

The dataset contains 37 numeric and 43 non-numeric independent variables, although some numeric fields should ultimately be considered as factors, e.g., month sold or overall quality ranking. In general, the data is well-populated with relatively few missing values.

In my research of the data, I identified four data fields - mainly involving home square footage - that were outliers. There was overlap across these fields: implementing my outlier removal code removed three observations. This aligns closely with the recommendation of the original author, that five data elements should be removed prior to sharing with students (De Cock).

As a preliminary dive into the data, I looked at the relationship between sale price, overall home quality (1-10 score), and square footage. As expected, the higher-quality home, the higher the sale price. Further, larger homes tended to be higher quality. That said, the collinearity between quality and size appears to reveal quality to be a stronger driver of sale price. Note that the neighborhood within Ames also appears to be an important driver of home sale price, which is an intuitive conclusion.

Digging into the year and month sold showed surprisingly little evidence of the initial hypothesis that post-crash sale prices would be notably lower than pre-crash (e.g., 2010 sales vs 2006 sales). That said, there is apparent month-level seasonality, and some variability when looking at sale price for each year-month over the observation window.

Feature Creation

I enriched the data by creating a metric of the total number of bathrooms in the home, summing the four metrics of full vs half baths in the basement vs aboveground. There is an apparent relationship where more total bathrooms are correlated with higher sale price, but the "n" becomes uncomfortably small at the higher end, which points to a "4+ bathrooms" category likely being more appropriate.

References

De Cock, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” *Journal of Statistics Education* 19, no. 3 (2011).

“House Prices - Advanced Regression Techniques.” Kaggle. Accessed January 9, 2022. https://www.kaggle.com/c/house-prices-advanced-regression-techniques/.

Indiviglio, Daniel. “After an Ugly 2010, the Housing Market Won't Look Much Better in 2011.” The Atlantic. Atlantic Media Company, September 6, 2013. https://www.theatlantic.com/business/archive/2011/01/after-an-ugly-2010-the-housing-market-wont-look-much-better-in-2011/69009/.