Michael A Moran

Statistics for Capstone Project

8 April 2018

Statistical Analysis of Housing Features

After the data wrangling steps were finished, I began to look at the data for features to which the sale price appeared highly sensitive. Using base R, I began to evaluate the sale price data by looking and the quartiles, mean, variance, and standard deviation (these were $34900, $129975, $163000, $214000, $755000; $180921; 6311111264; and $79443, respectively). I initially began to create data visualizations also using base R, but the coding required proved time consuming and the results were visually unimpressive. At Jeff’s suggestion I installed Tableau and uploaded the cleaned data file into a workbook. By dragging and dropping variables and easily switching between graphics types, I was able to generate data visualizations in Tableau in a fraction of the time, with much more striking visuals.

I started out by again narrowing the list of what I thought might be relevant variables to the final price down to twelve and generating visualizations for each of the twelve variables. These variables were MSSubClass, the type of dwelling involved; FullBath, the number of full bathrooms; BedroomAbvGr, the number of bedrooms; GarargeType, the location of the garage; BldgType, another dwelling classification; SaleCond, the condition of the sale; OverallCond, the overall condition rating; LotConfig, the lot configuration; BsmtCond, the overall condition of the basement; RoofStyle, the type of roof attached; Condition1, the proximity to a main road or artery; and GrLivArea, the total square footage above ground. Additionally, I generated visualizations for the percentage of houses in the dataset of each bedroom number and the median house price by Neighborhood at Jeff’s suggestion.

After exporting a .pdf of the visualizations and uploading the file to GitHub, I reviewed the outputs with Jeff as well my conclusions from the data. We determined from the graphic analysis that BldgType, LotConfig, RoofStyle, and Condition1 were not significantly related to the sale price. These outputs showed distributions that were nearly identical but for the prevalence of one of the type over the others. This indicated the popularity of a style but did not indicate a relationship to the sale price. Additionally, while the sale condition showed a small relationship to the sale price, it was a sensitivity that corresponded to only a fraction of a standard deviation of the sale price and ultimately indicated that houses generally sold for what they were worth on the open market, regardless of the circumstances surrounding the sale. A relationship to the sale price was observed for the remaining variables.

Strong linear relationships were observed between sale price and the numbers of bathrooms, the total square footage, and the overall condition of the house. On the graph of sale price as a function of square footage, it can be observed that the data distribution gets looser the higher the square footage climbs, indicating that at higher square footages there is a greater variability of sale price. The number of bedrooms exhibited a strong linear relationship from two bedrooms to four, then decreases for five and six bedrooms before jumping again for eight bedrooms. In concert with the visualization of the percentages of each number of bedrooms, I determined that since most of the houses fall in the range up to four bedrooms, where the trend is strong, I would include the number of bedrooms in the regression model.

Associations were also observed in the distribution of sale prices for each of the types of garage and basement condition variables, as well as in the distribution of the sale price by MSSubClass, which describes the building with respect to both the type of structure (single fam, townhouse, duplex, etc.) and when it was constructed (pre or post 1946). The garage types also showed greater variability of the sale price associated with These associations were weaker than the other already discussed but were deemed significant enough for the variable to be included in the initial regression analysis. Lastly, I considered the Neighborhood variable, which showed a high sensitivity to the median sale price. After conferring with Jeff about this, we decided that I should include the Neighborhood in my initial regression analysis and would need to go back to create dummy Neighborhood variables.

After considering the statistical analysis generated, I decided to start with eight variables in my regression analysis with the goal of eliminating three. I selected the number of bathrooms, the number of bedrooms, the total square footage, the overall condition of the house, the neighborhood, the type of garage, the basement condition, and MSSubClass. These variables are both intuitive as affecting and graphically indicative of a relationship with the sale price.