Lauren Thompson

DSC530

Data Exploration and Analysis

Summary

The goal with this exploratory data search of the housing market was to identify what variables can be used to predict the price a house sales for and find trends over the years. Questions surfaced as the exploratory portion took place such as how the square footage of the lot itself plays a role in pricing, and if so, does it play a more significant role than the square footage of the house? How do the number of bedrooms and type of bathroom effect the price? The overall goal was to explore how sale prices fluctuates depending on variables commonly looked at when pricing a property.

Location, location, location. Many of the tests performed weren`t significant or were not a good fit of a model. If these test and models were to be run again after isolating each state, they may produce more promising results. By not having any filter or separation of region it opened the doors for outliers and convoluted data. For example, a 3-bedroom 2.5 bath house being sold in Maryland is priced much higher than a 3-bedroom 2.5 bath in Texas. Or even switching zip codes could alter the price significantly. For example, a house on the outskirts of Frederick County Maryland will be much less than one in the heart of Frederick County.

Another variable that could have been helpful is asking price as well as how much the house is appraised for, how many houses were on the market, and how long the house has been on the market. The true appraisal value could shed light on patterns of asking prices depending on how many houses are on the market and how long houses are staying on the market. Essentially these added variables would identify supply and demand ratios.

A beneficial analysis that was not performed in this exploratory analysis was a heat map. Having a heat map with different layers to it such as the variables under investigation could highlight areas of interest and ways to separate the data. Organizing the data differently would have helped even the playing field. Having subsets of housing with the same specs and similar lot sizes could have given more information about how location effects prices.

An assumption made from the start of this data exploration and analysis was that the database was a good and appropriate sample and population size to represent the United States as a whole. If this project were to be done again, I would break the data out by state and potentially zip code and see how much data was there. Then a scale to size comparison would be made to ensure the sample size is truly representing the majority. For example, it would be alarming if there were only 50 homes old in Texas in a given year and 1000 sold in Maryland.

Final thoughts, all houses are not created equal.