Nihar Varanasi
04/13/2021

Exploratory Data Analysis with Housing in Ames, IA

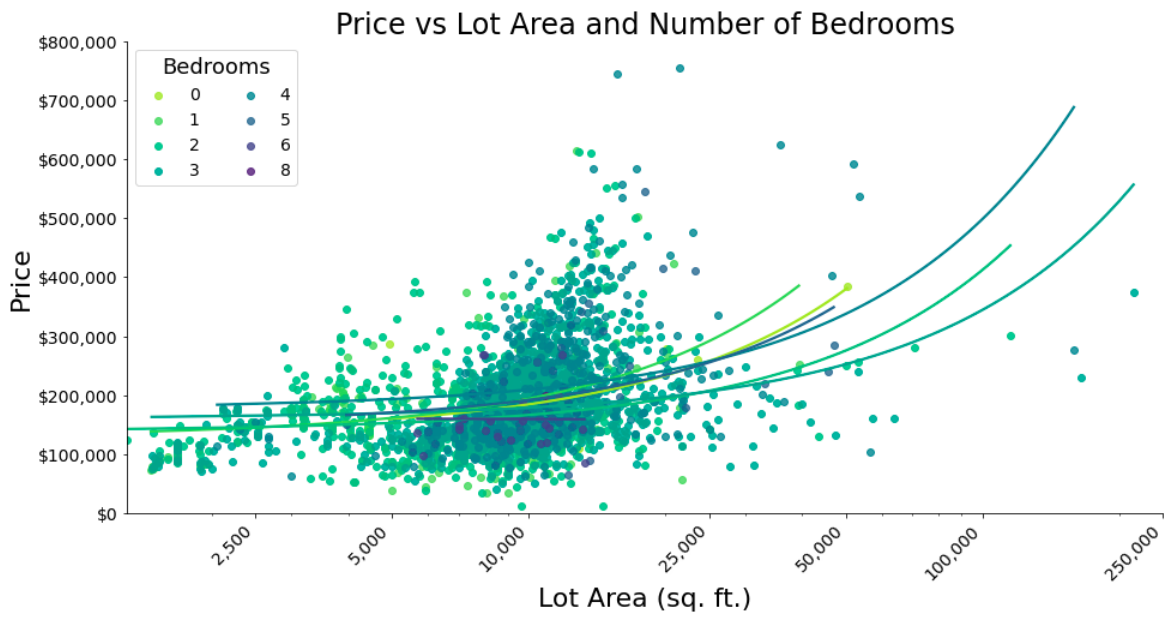
Dataset Description and Exploration Plan:

For this assignment, I am using the Ames, Iowa dataset from Kaggle. This dataset contains 2930 observations and 82 variables related to house sales in Ames, Iowa. The variables are related to characteristics and measurements of the house, its location, and its purchase. I am going to select a few variables that are noteworthy and compare relationships between them and to *Sale Price*. The variables are varied between numerical and categorical, so I plan to choose interesting variables that would be of common interest that also contain a minimal number of missing values.

Data Cleaning and Feature Engineering:

I am looking at *Sale Price*, *Lot Area*, *Number of Bedrooms*, *Neighborhood*, *Year Built*, and *Overall Quality*. After careful analysis of the structure of data, there are some missing values for this subset, but the data is majority complete and in good quality. There is also no feature engineering being implemented in the dataset.

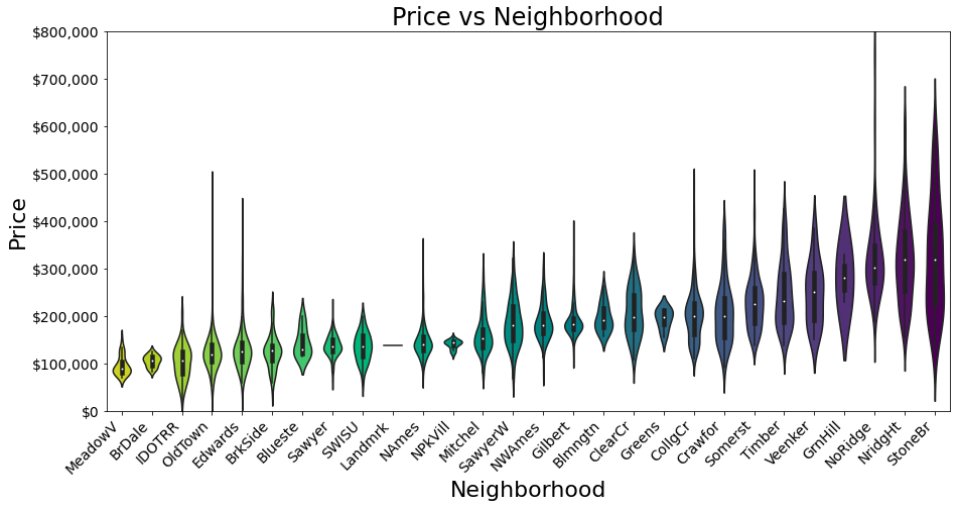
Exploration 1:



Number Of Bedrooms	Count Of Houses
0	8
1	112
2	743
3	1579
4	400
5	48
6	21
8	1

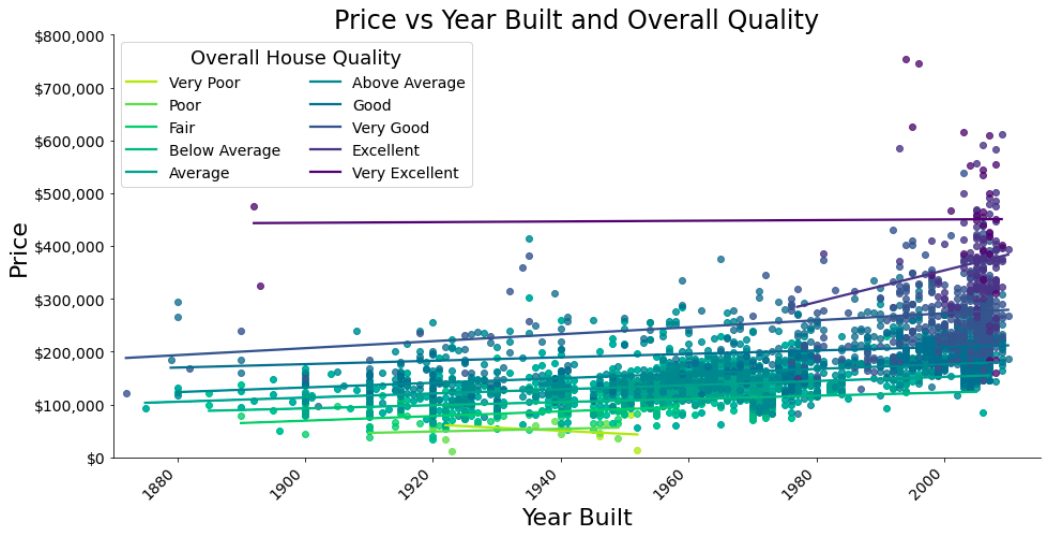
For the first exploration, *Price* is compared to *Lot Area* and Number of *Bedrooms*. The variable, *Lot Area*, had a considerable right skew, so the x-axis was log scaled. *Number of Bedrooms* alone does not indicate a clear relationship with price. However, all regression lines for *Number of Bedrooms* from 0 to 6 show a positive correlation of *Lot Area* with *Price*. As noted above in the table, there are only a few houses with 6 or 8 bedrooms.

Exploration 2:



For the second exploration, *Price* is compared to *Neighborhood*. *Neighborhood* is represented by violin plots, sort in ascending order by median *Price*. There is much overlap in *Price*, as there is significant and varying spread of house values within each *Neighborhood*. Overall, based on the violin plots, there seems to be some correlation that warrants a further investigation.

Exploration 3:



For the third exploration shown above, *Price* is compared to *Year Built* and *Overall Quality*. Comparing these three together is quite telling. Houses built earlier are of overall lower quality and lower selling price, as seen with the yellow, green, and cyan data points in the bottom of the graph. From 1990 - 2010, there is a sharp increase in quality and selling price, as seen with the blue, indigo, and purple data points on the far right of the graph. The regression lines for *Overall Quality* show a positive relationship for *Year Built* and *Sale Price*. Therefore, there is very little intersection of these lines, they are clearly separated. The exception is Very Poor and Poor at the bottom center of the graph.

Hypothesis Testing:

Based on the data, I wanted to look at three different tests. The first hypothesis test was to see if an increase in *Lot Area* will increase *Sale Price*. The second hypothesis test was to see if the homes located in Green Hills, Northridge, Northridge Heights, and Stone Brook have a greater *Sale Price* than homes located in other *Neighborhoods.* Lastly, the third hypothesis was to see if an increase in *Overall Quality* will increase *Sale Price*. After going through these three tests, I am going to conduct the experiment for the third hypothesis.

Significance Testing:

H_o: *Overall Quality* does not affect *Sale Price*

H_a: *Overall Quality* does affect *Sales Price*

After running the test, the R^2 value is 0.639 and the p-value is <0.0001. Based on the R^2 value, 63.9% of the variance in *Sale Price* can be explained by *Overall Condition*. Based on the

p-value, this is significant at the 95% confidence level. Therefore, we can reject the null hypothesis and say that the *Overall Quality* does affect the *Sales Price*.

Data Summary and Future Analysis:

Although there are some missing values, the vast majority of this data is complete and of good quality. Latitude and Longitude of houses, that is, there specific locations, may show to be a better indicator of price than *Neighborhood*. It is important to stress that this model is specific to Ames, IA, and would vary greatly across the US. Although this data set of nearly 3,000 observations is satisfactory for a city, a much larger sample would be necessary for the US.

For the future, determining other variables with a high level of correlation, doing a multiple regression to explain more variance in *Sale Price* to achieve significant results would be the next steps in this analysis.