Notebook

August 1, 2019

0.0.1 Question 1a

Based on the plot above, what can be said about the relationship between the houses' sale prices and their neighborhoods?

All the saleprices are around 200000, no matter how many neighborhoods or what is the neighborhood. So there is no relationship between the houses' sale prices and their neighborhoods.

0.0.2 Question 3a

Although the firepalce quality variable that we explored in Question 2 has six categories, only five of these categories' indicator variables are inclulded in our model. Is this a mistake, or is it done intentionally? Why?

It is done intentionally as the drop_first is set to be True. The reason is that one dummy variable is the linear combination of the others.

0.0.3 Question 3d

Compare the predictive accuracy of this model to that of the model that you derived in Homework 5. Is the new model a better predictor of housing prices in Ames? If so, are the gains in accuracy significantly larger? Assume that the training and testing sets used to in Homework 5 are identical to the ones used in this homework.

Write your answer here, replacing this text.

0.1 Question 5: EDA for Feature Selection

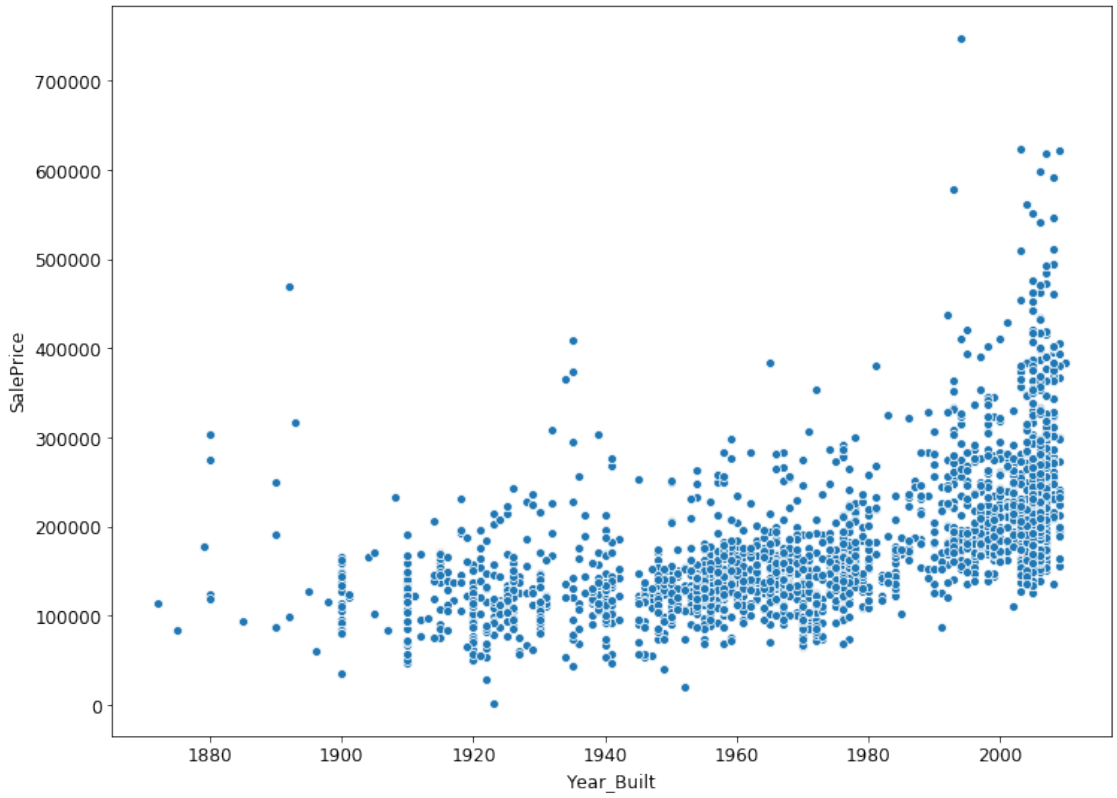
In the following question, explain a choice you made in designing your custom linear model in Question 4. First, make a plot to show something interesting about the data. Then explain your findings from the plot, and describe how these findings motivated a change to your model.

0.1.1 Question 5a

In the cell below, create a visualization that shows something interesting about the dataset.

```
In [31]: # Code for visualization goes here
         sns.scatterplot(training_data['Year_Built'],training_data['SalePrice'])
```

```
Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x7fcc043e9518>
```



0.1.2 Question 5b

Explain any conclusions you draw from the plot above, and describe how these conclusions affected the design of your model. After creating the plot, did you add/remove certain features from your model, or did you perform some other type of feature engineering? How significantly did these changes affect your rmse?

From this plot, we know that the 'sale price' is positive correlated to the 'year built'. I will add this feature to my model. Both of training rmse and testing rmse decrease about 1000 after adding this feature