Notebook

April 8, 2020

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?

House prices are mostly between 100000 and 300000. It cannot be said that the price is set according to the neighbors, but in some cases, it seems to have a great influence. For example, in the case of 'NridgHt' and 'StoneBr', the median is ranked above 300,000, and the price of the top 25 pros exceeds 400000. The number of counts is 112 and 28, respectively, which cannot be ignored, and the price of the lower 25% is also higher than other neighbors.

0.0.2 Question 3a

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

I think the above expression is fine. The removed variable is "average", and we can consider it as "0". In this case, it can be said that a step above "average" has a positive value, and a step below it has a negative value.

0.1 Question 5: EDA for Feature Selection

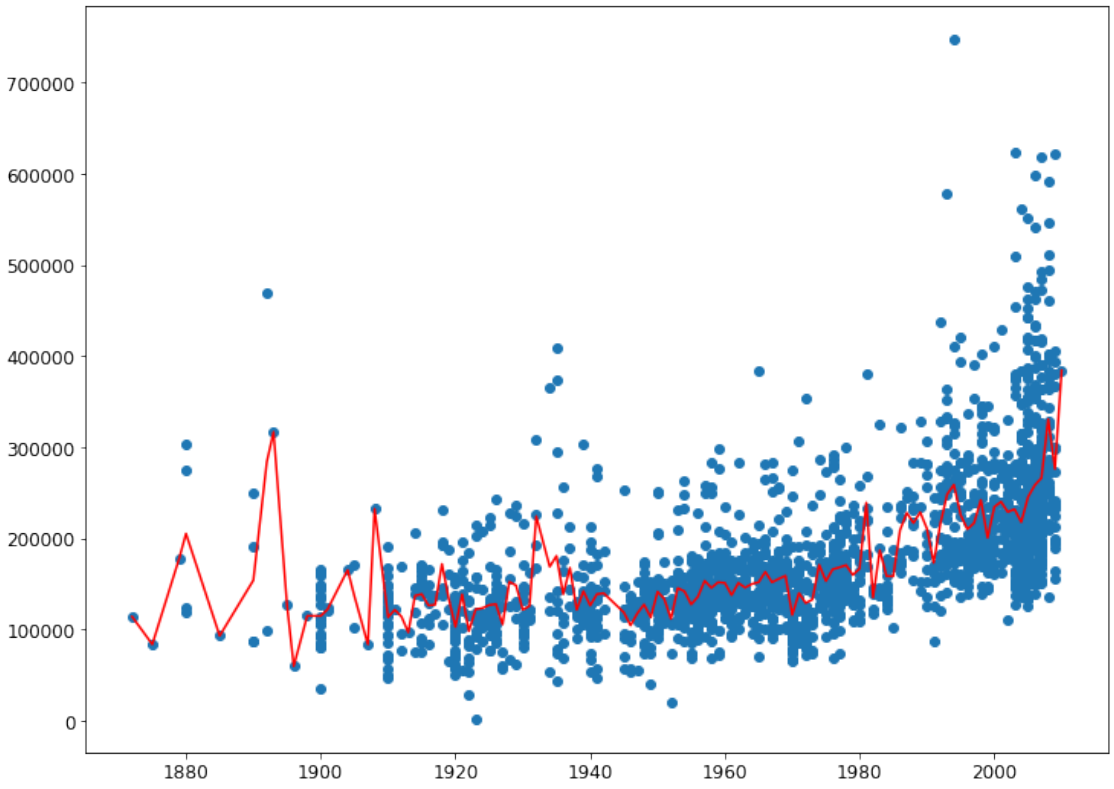
In the following question, explain a choice you made in designing your custom linear model in Question 4. First, make a plot to show something interesting about the data. Then explain your findings from the plot, and describe how these findings motivated a change to your model.

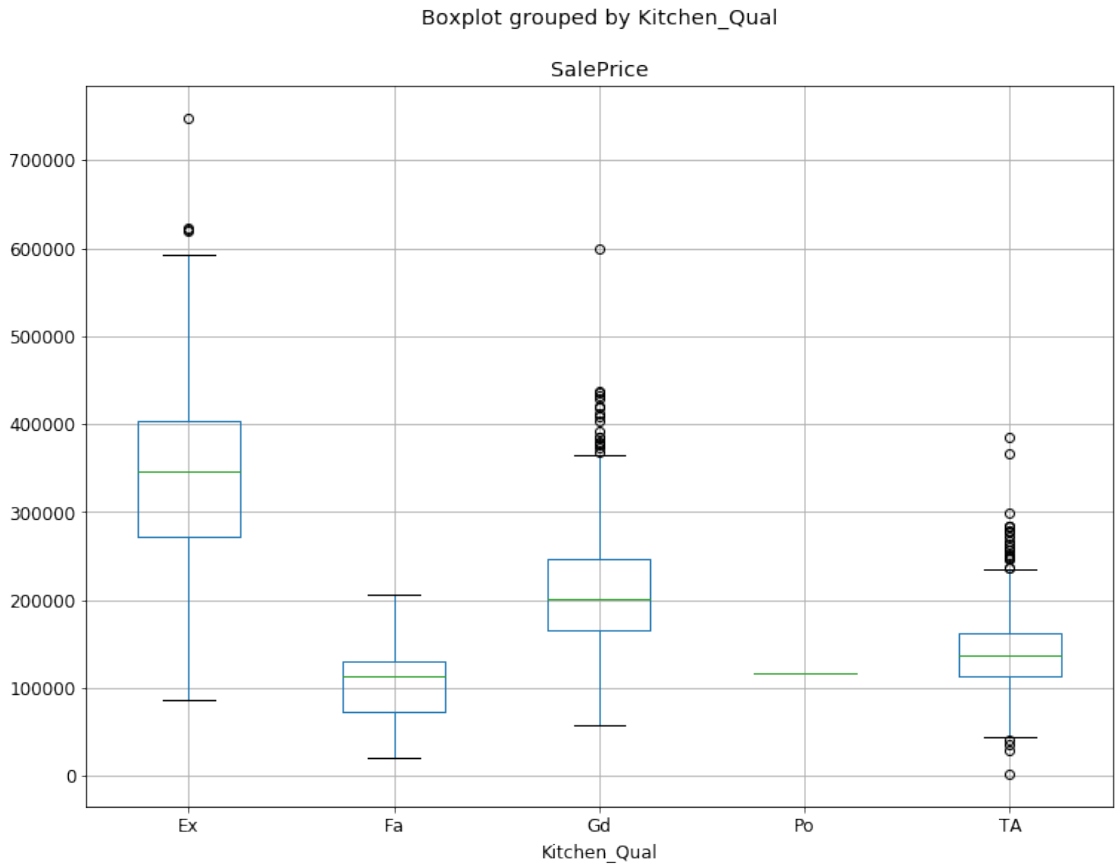
0.1.1 Question 5a

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

```
In [31]: # Code for visualization goes here
         sz1 = training_data.groupby('Year_Built')['SalePrice'].mean();
         plt.scatter(training_data['Year_Built'], training_data['SalePrice'])
         plt.plot(sz1, 'r')
         training_data.boxplot('SalePrice', 'Kitchen_Qual')
```

```
Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x7efef5def898>
```




Boxplot grouped by Kitchen_Qual
SalePrice
700000
600000
500000
400000
300000
200000
100000
0
Ex
Fa
Gd
Po
TA
Kitchen_Qual

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?

I set 'Kitchen_Qual' and 'Year_Built' as variables to add to my model. We can see that in the case of 'Kitchen Quality', a building with a better rating than 'average (TA)' forms a higher price range. As a result of Scatter plot, you can see the points rising upward as the building is built more recent. Although there are differences from year to year, looking at the overall line, the average building price for each year seems to rise upwards. By putting these two variables into the model, I was able to make better predictions.