Notebook

April 6, 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?

On average, there are less houses in neighborhoods with higher sale prices.

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?

This was done intentionally since having 6 category indicator variables is redundant. By taking out Ta (average), you will know it has to be average is it is not any of the others. Having 5 will also allow us to have a full rank matrix, which allows us to calculate the least squares regression. Having 6 category indicators will make it linearly dependent.

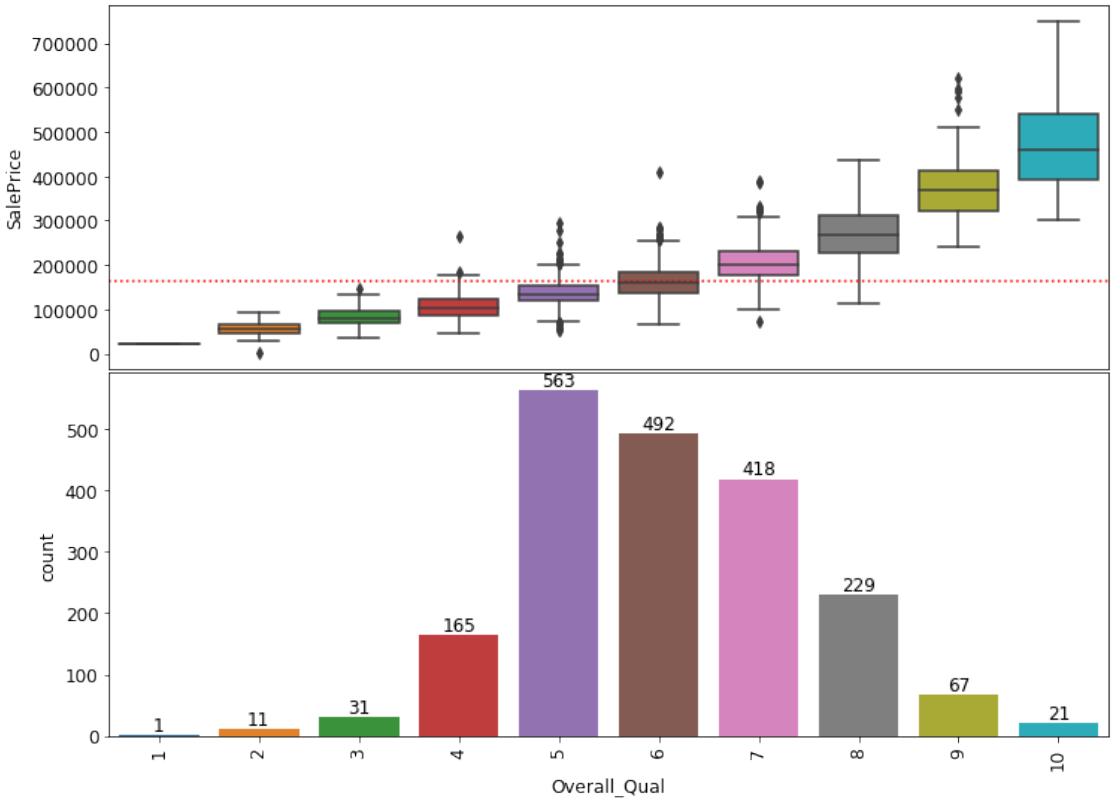
0.1 Question 5: EDA for Feature Selection

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 [118]:
```
# Code for visualization goes here
fig, axs = plt.subplots(nrows=2)

sns.boxplot(
    x='Overall_Qual',
    y='SalePrice',
    data=training_data.sort_values('Neighborhood'),
    ax=axs[0]
)

sns.countplot(
    x='Overall_Qual',
    data=training_data.sort_values('Neighborhood'),
    ax=axs[1]
)

# Draw median price
axs[0].axhline(
    y=training_data['SalePrice'].median(),
    color='red',
    linestyle='dotted'
)

# Label the bars with counts
for patch in axs[1].patches:
    x = patch.get_bbox().get_points()[:, 0]
    y = patch.get_bbox().get_points()[1, 1]
    axs[1].annotate(f'{int(y)}', (x.mean(), y), ha='center', va='bottom')

# Format x-axes
axs[1].set_xticklabels(axs[1].xaxis.get_majorticklabels(), rotation=90)
axs[0].xaxis.set_visible(False)

# Narrow the gap between the plots
plt.subplots_adjust(hspace=0.01)
```


SalePrice
700000
600000
500000
400000
300000
200000
100000
0
count
563
492
418
229
165
67
31
21
11
1
1
2
3
4
5
6
7
8
9
10
Overall_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?

Based on the plot above, the better the overall quality of the house, the higher the sale price is, and the lower the quality of the house is, the lower the sale price is. Most houses have an average quality of 5 and above, with 6 being at the median sale price. Adding this feature dramatically decreased my rmse. After adding overall quality as a feature, I removed fireplace quality, but this raised my rmse slightly so I decided to put it back.