Notebook

July 30, 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?

1. There is no distinct relationship between the houses'sale prices and their neighborhoods. For example, for those neighborhoods with high counts, some of their houses have relatively high prices, such as NAmes, while the others have very low price, such as OldTown.
2. The counts are quite different among all the neighborhood, where NAmes has a count of 299, while Greens has only a count of 2. Therefore, the mean house price in NAmes is much lower than the mean price in Greens.

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's done intentionally. Because the sum of five categories'indicator variables will always be 1, so the five columns will always be linearly dependent, which is a problem for invertibility . Also, it will affect how we interpret the model weight.

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.

The new model is a better predictor of housing prices in Ames. In hw5, the pair (training error, test error) is (46710.597505875856, 46146.64265682625). The new pair (training error, test error) is (40491.84911146645, 38754.86068184426), where both errors decline sharply. Therefore, the gains in accuracy significantly larger.

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 [278]:
```
# Code for visualization goes here
sns.boxplot(
    x='KitchenQual',
    y='SalePrice',
    data=training_data
)
```

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?

Write your answer here, replacing this text.