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?

The houses' sale prices and their neighborhoods are quite related.

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?

Intentionally. We need the X matrix to be full ranked which means that rows of X has to be independent. If we include all of the 6 catogories then the 6th catogories can be calculated from the first 5, rendering rows of X to be 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

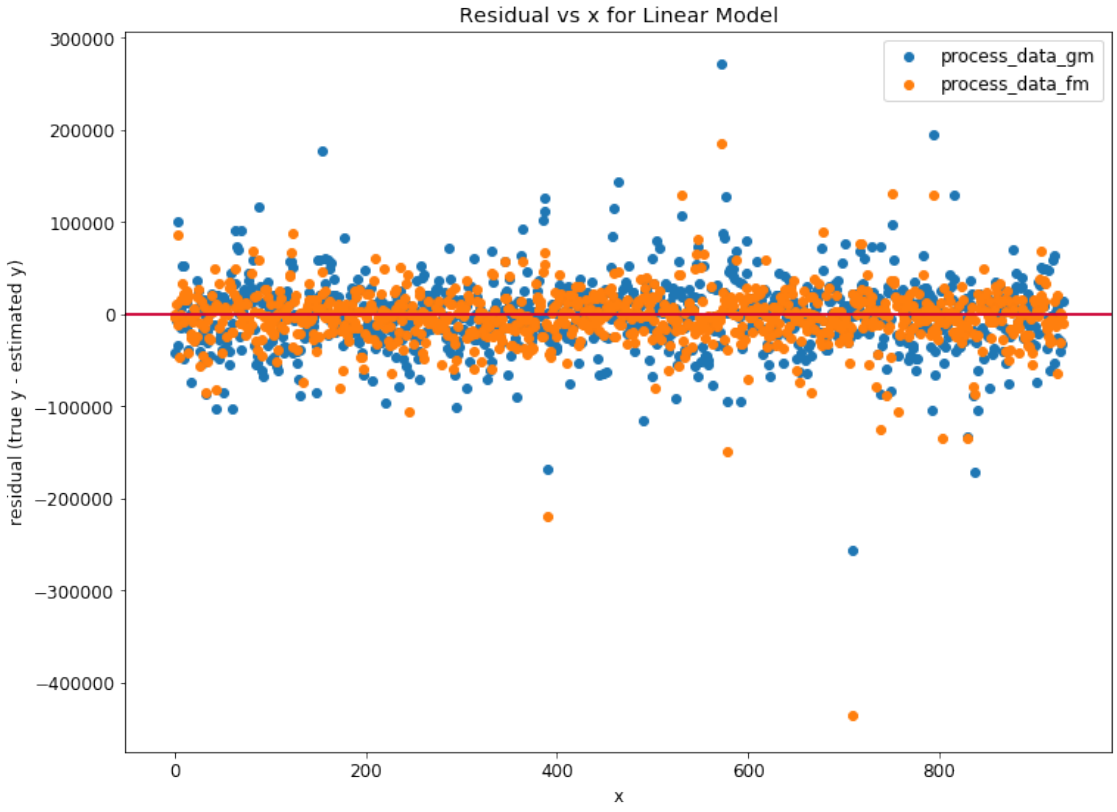
In the cell below, create a visualization that shows something interesting about the dataset.

In [31]:
```
# Code for visualization goes here
test_data = pd.read_csv('ames_test_cleaned.csv')

X_test, y_test = process_data_gm(test_data)
X_test = X_test.fillna(0)
plt.scatter(np.arange(len(X_test)), y_test - linear_model.predict(X_test), label='process_data_
plt.xlabel('x')
plt.ylabel('residual (true y - estimated y)')
plt.title('Residual vs x for Linear Model')
plt.axhline(y = 0, color='b')

X_test, y_test = process_data_fm(test_data)
plt.scatter(np.arange(len(X_test)), y_test - final_model.predict(X_test), label='process_data_f
plt.xlabel('x')
plt.ylabel('residual (true y - estimated y)')
plt.title('Residual vs x for Linear Model')
plt.axhline(y = 0, color='r')

plt.legend();
```



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?

Looking into the process_data_gm we can conclude that the guiding processed model is good as points are randomly scattered around the line y = 0. So instead of tweaking the features we have already included in the guiding model I decided to add all possible numerical features into the final processing model. Result plot shows that such change significantly lower the rmse.