Housing Price Predictions

By: Daniel Gallo

Agenda

- Data Processing
- Exploratory Data Analysis
- Machine Learning
- Results

Data Processing

Data Overview

- Shape of the Data: 2930 rows, 82 columns.
- Many missing values
- Good amount of outliers
- In the end, only had to drop 4 rows, and dropped no columns.
- Added a couple of columns

Missing Values

- Many of the missing values, were not missing values at all. They had a meaning.
- Using the data dictionary, was able to fill in the missing values with the necessary meaning.
- Was able to fill in the missing values in all but 4 rows, which were dropped.

Outliers

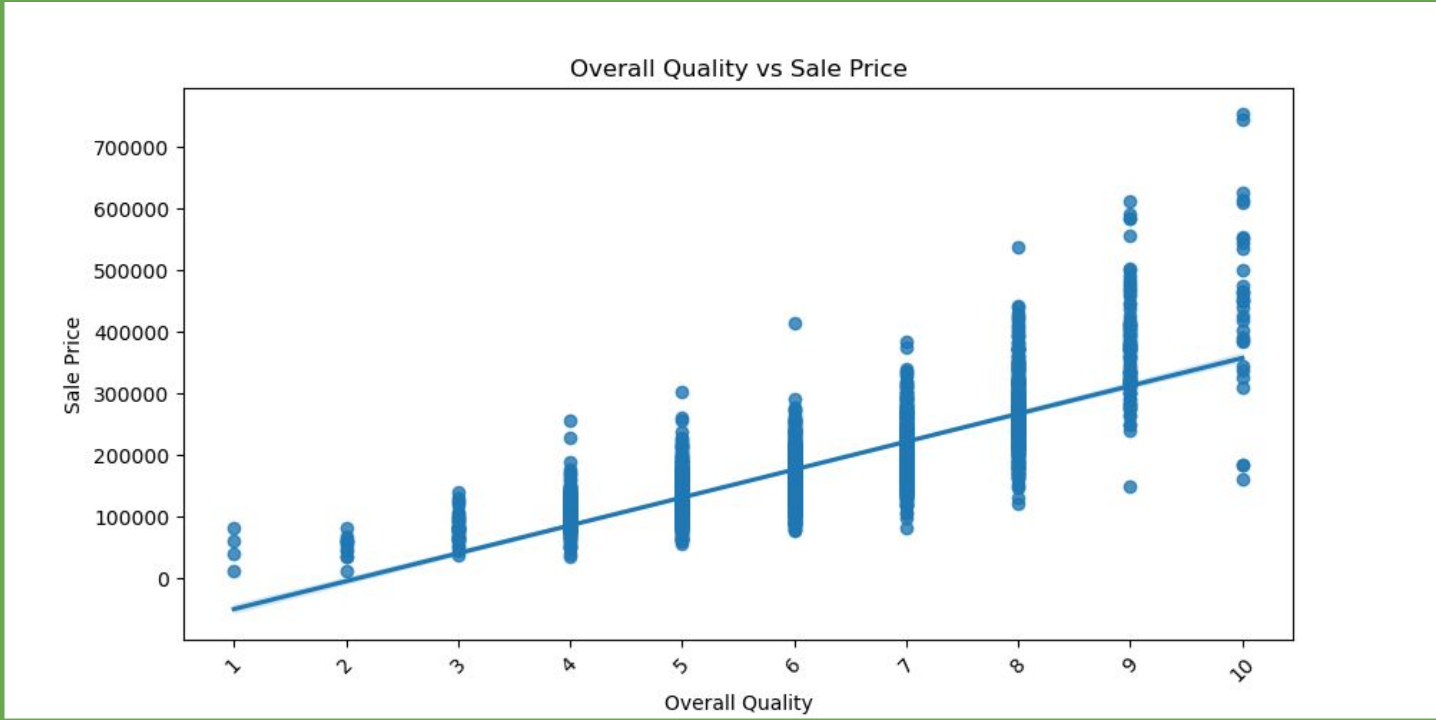
- With the exception of one, all outliers remained in the dataset, so that the model could learn that some houses may have extreme values.
- Outlier removed was a year far in the future, so was replaced with an NaN, then filled with the median for other missing years.
- Median was used to fill it in so as to have a minimal effect on the algorithm.

Feature Engineering

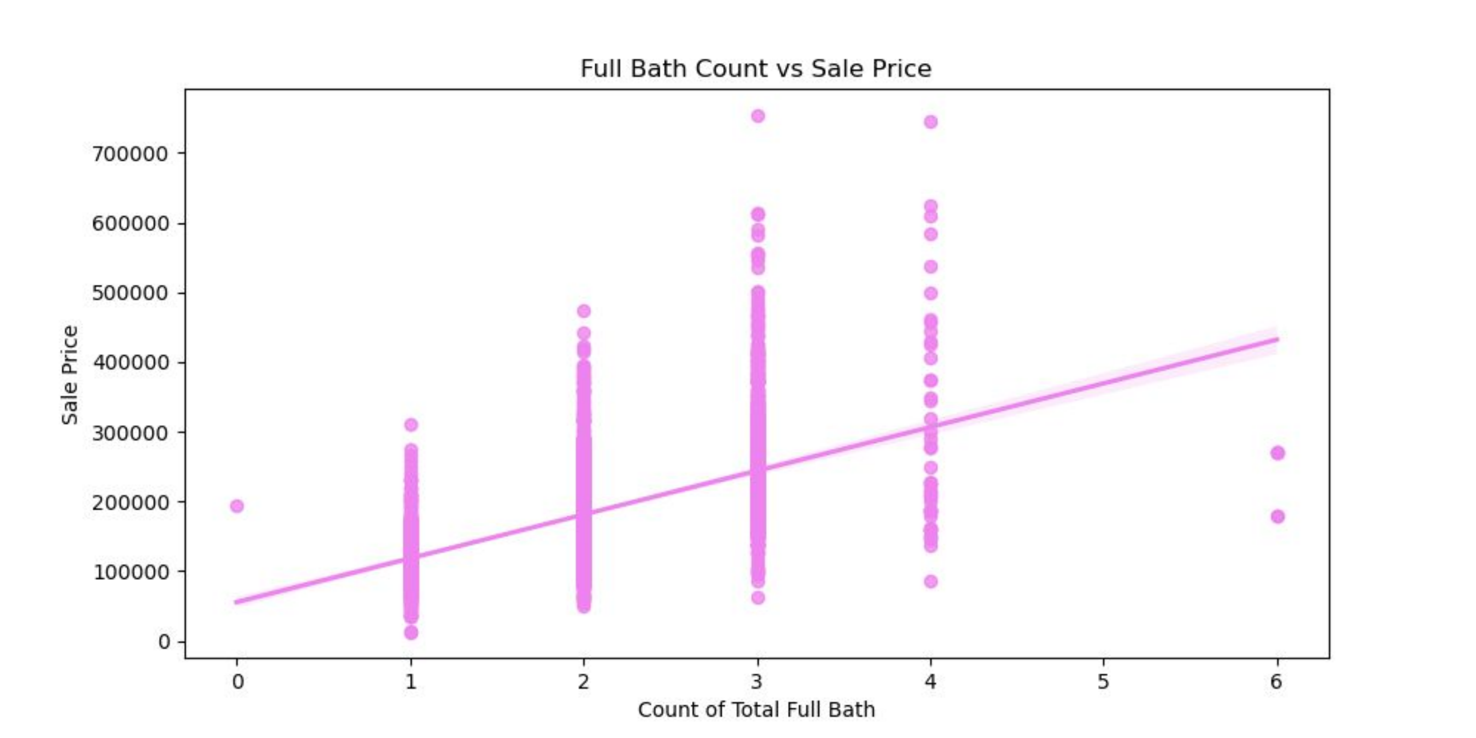
- Many of the columns, namely the ordinal and nominal categories, were all converted to categorical column types.
- The ordinal columns were then further manipulated to have numerical codes for each category, depending on the column.
- This will allow the columns to be passed into the machine learning model, as it can only use numerical values.
- Also combined the ‘Full Bath’ and ‘Bsmt Full Bath’ columns into ‘Total Full Bath’, to have a full view into how many bathrooms there are. Same with the half baths.

Exploratory Data Analysis

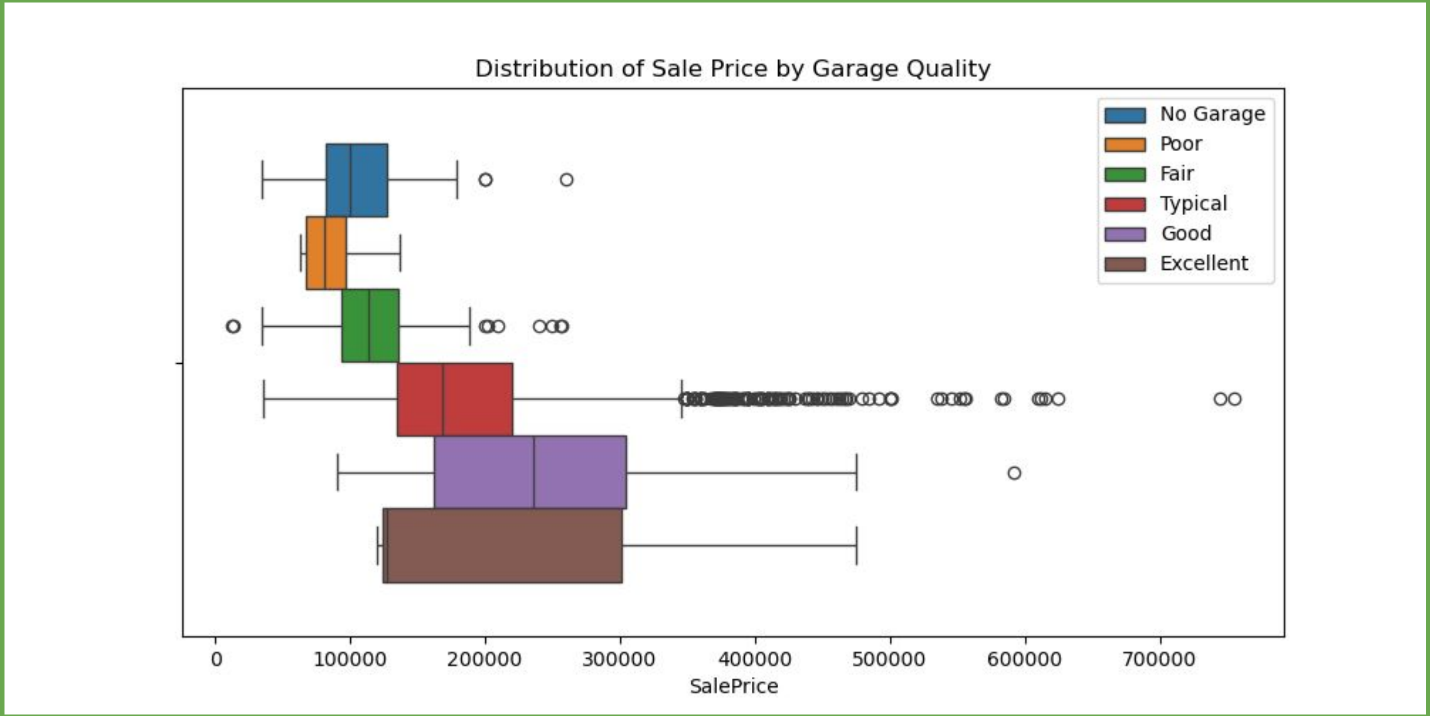
How does the overall quality of the house relate to the sale price?



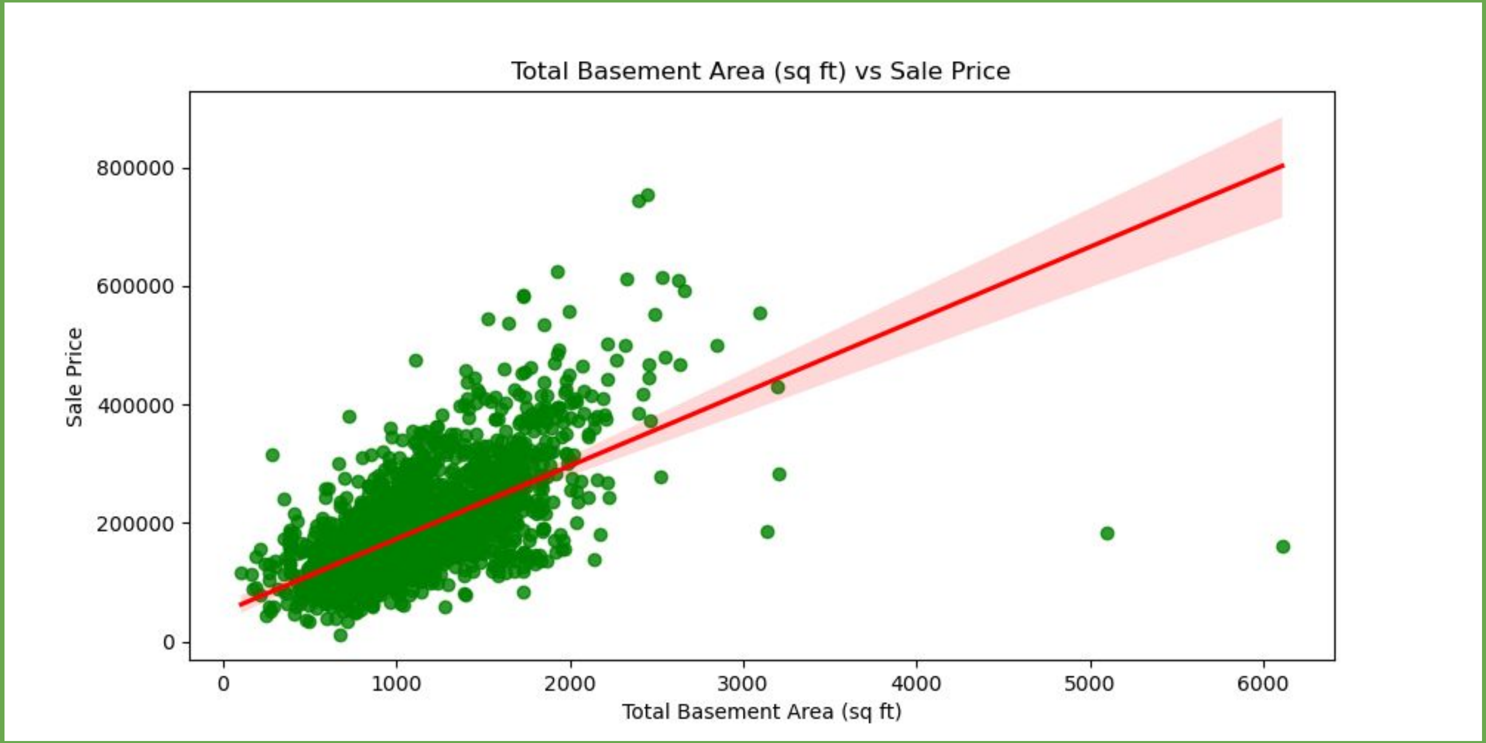
Does the number of full bathrooms affect the sale price?



Does garage quality make a difference in the distribution of sale price?



How does total basement area affect the sale price?



Machine Learning

Model Preprocessing

- To decide the features, I needed to look at the types of data I had.
- Not everything was numeric
- I used pd.get_dummies to convert nominal category columns to numeric
- I then kept a list of the numerical features and nominal features so that I don't need to make dummies for every column.

Feature Selection

- I used 26 features in total
- I used the condition and quality codes that were most relevant.
- I also used information about the size of the house/garage/basement, as well as the number of rooms in total.
- Nominal columns about the building type, the house style, and the neighborhood were also considerations.

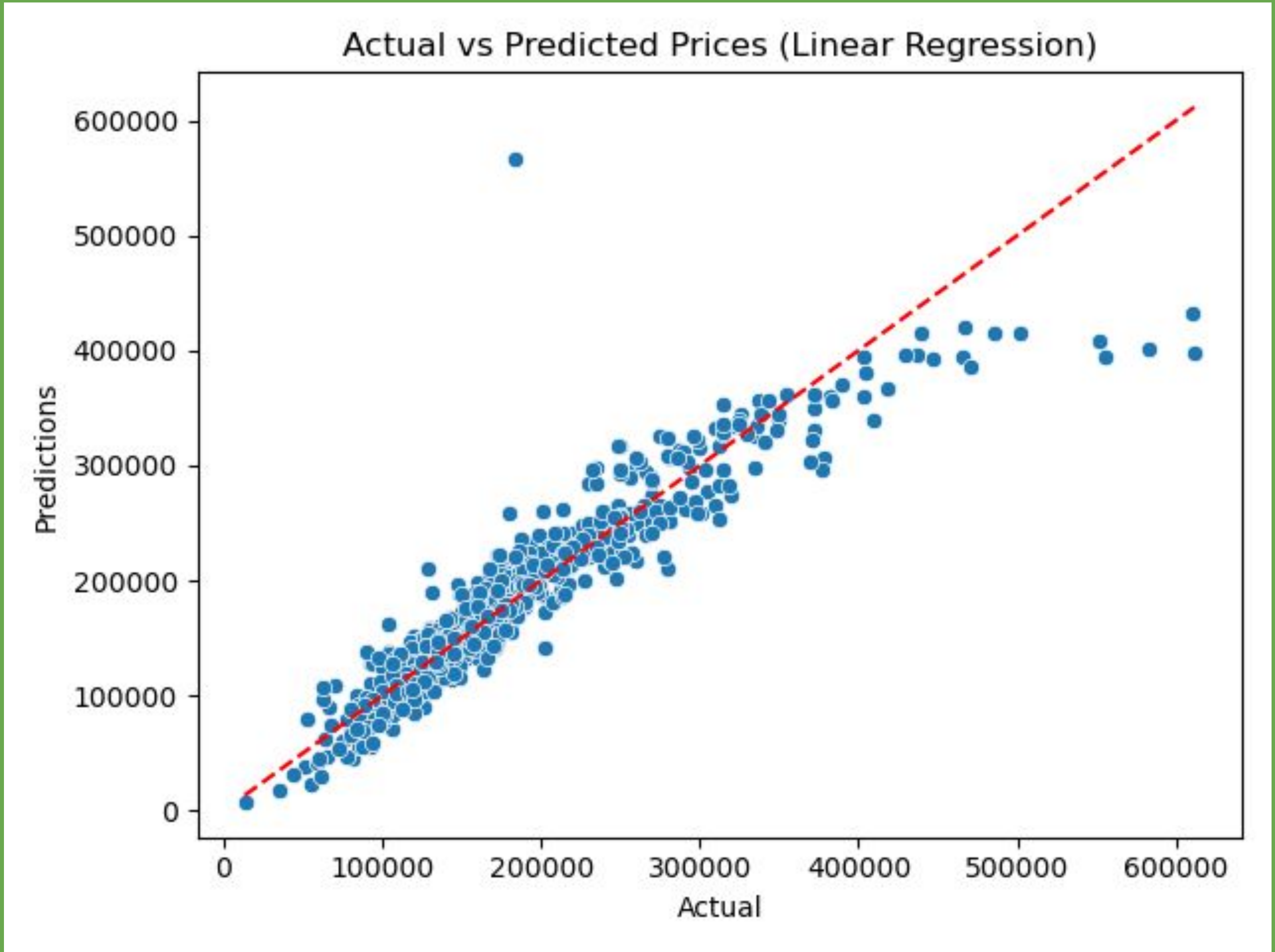
Model Training

- 75/25 split for training and testing sets.
- I used Linear Regression as the main algorithm for this dataset.
- After fitting the model, I used a custom function to output the R^2 score, as well as the Root Mean Squared Error (RMSE) score.

Results

Model Results

- R^2 score: ~86%
- RMSE ~$30,000
  - The mean house price in this data set was around $180,810. This would be an error around 16.6%.


Actual vs Predicted Prices (Linear Regression)
Predictions
Actual
0
100000
200000
300000
400000
500000
600000

Other Models Used

- I also used ridge and lasso regression to regularize the data, but this had minimal effect, and generated near identical results as the linear regression.
- I also used regression models like Random Forest Regressor, Linear SVR (Support Vector Regression), and Decision Tree Regressor.
- The Random Forest regressor achieved much higher R^2 scores and lower RMSE scores than the linear model.
- The LinearSVR model performed poorly on the data
- The decision tree did well with the training data, but not as well with the test data.


QUESTIONS