Ames Housing Price Predictions

By: Daniel Gallo

Agenda

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

Problem Statement

The goal of this project is to find what features have strong relationships with the sale price, and could be used to create a model to predict housing prices in the future.

Data Processing

Data Overview

- This dataset had 2930 individual houses, with 82 total features. Not every house has a value for each feature, so I had to fill in the missing data.

Missing Values & Outliers

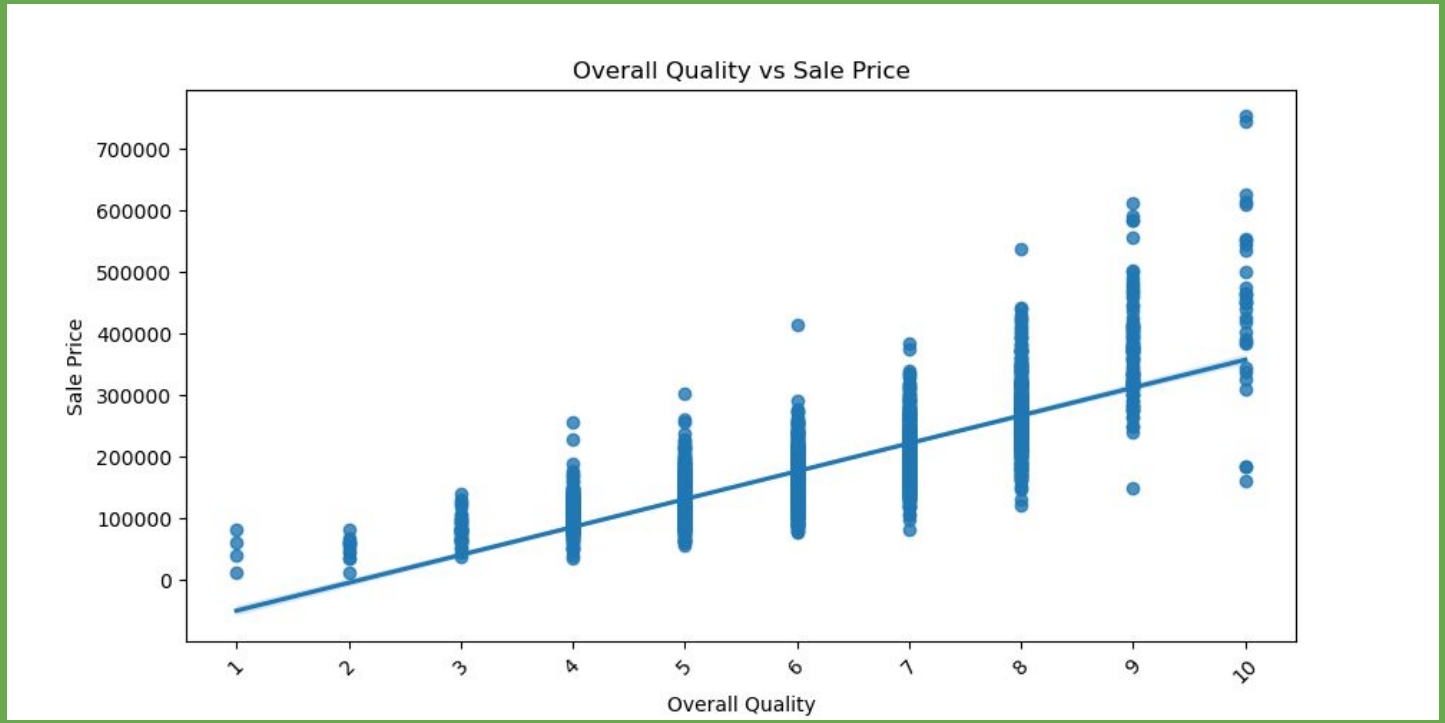
- This dataset contained many different missing values, which I was able to fill in through my research of the data.
- There were also many outliers, or extreme values, in the data, but I opted to keep them so my model could learn from it.

Feature Engineering

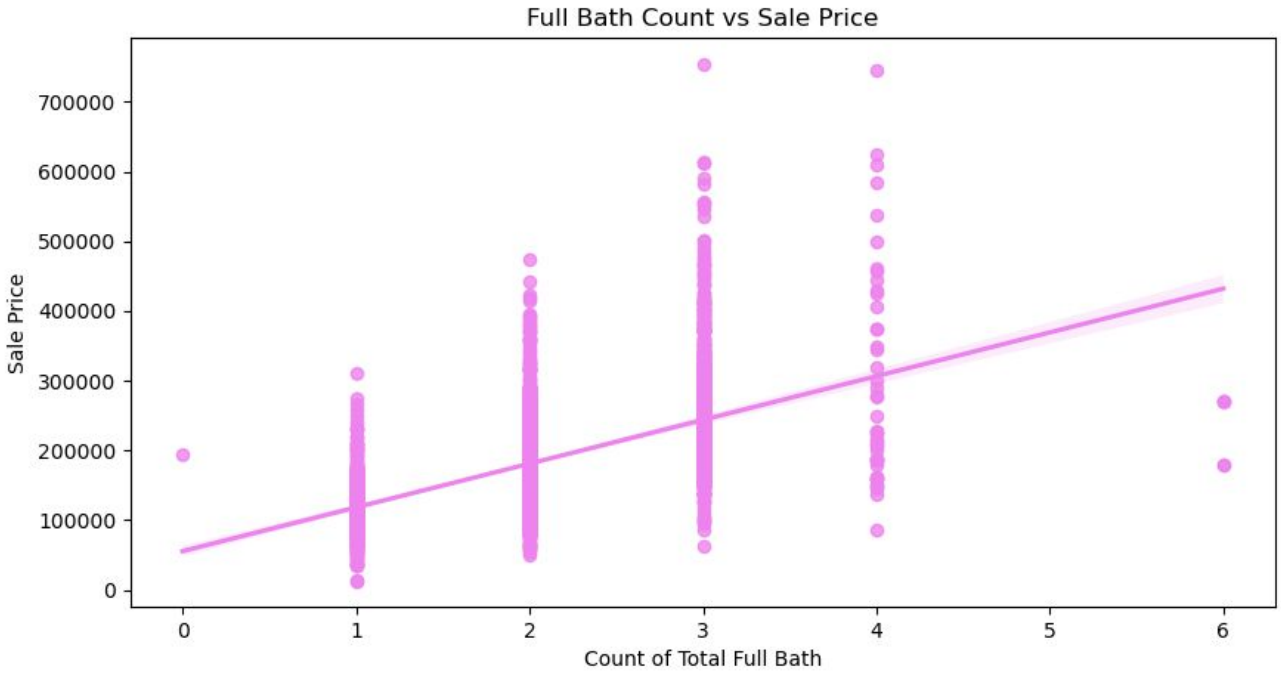
- Many of the columns, namely the ordinal and nominal categories, were all converted to categorical column types.
- The ordinal columns were then further manipulated to have numerical codes for each category, depending on the column.
- This will allow the columns to be passed into the machine learning 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.

EDA

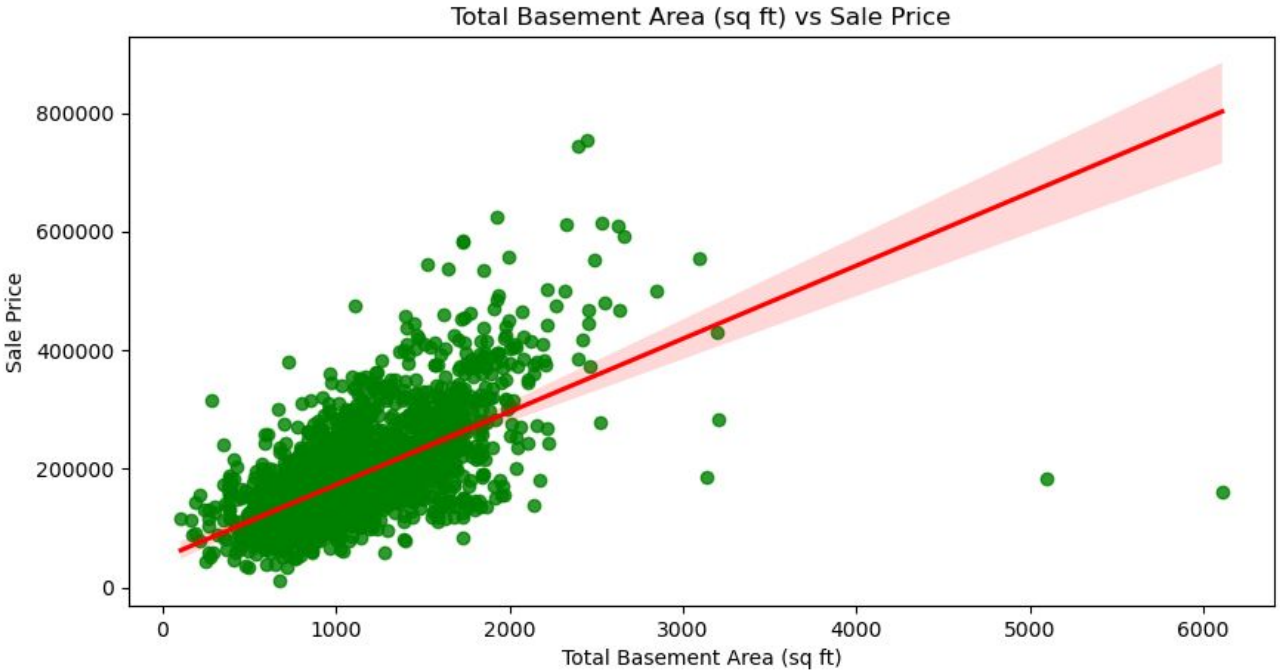
How does the overall quality of the house relate to the sale price?



Does the number of full bathrooms affect the 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 converted nominal category columns to numeric format with dummy variables
- 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.

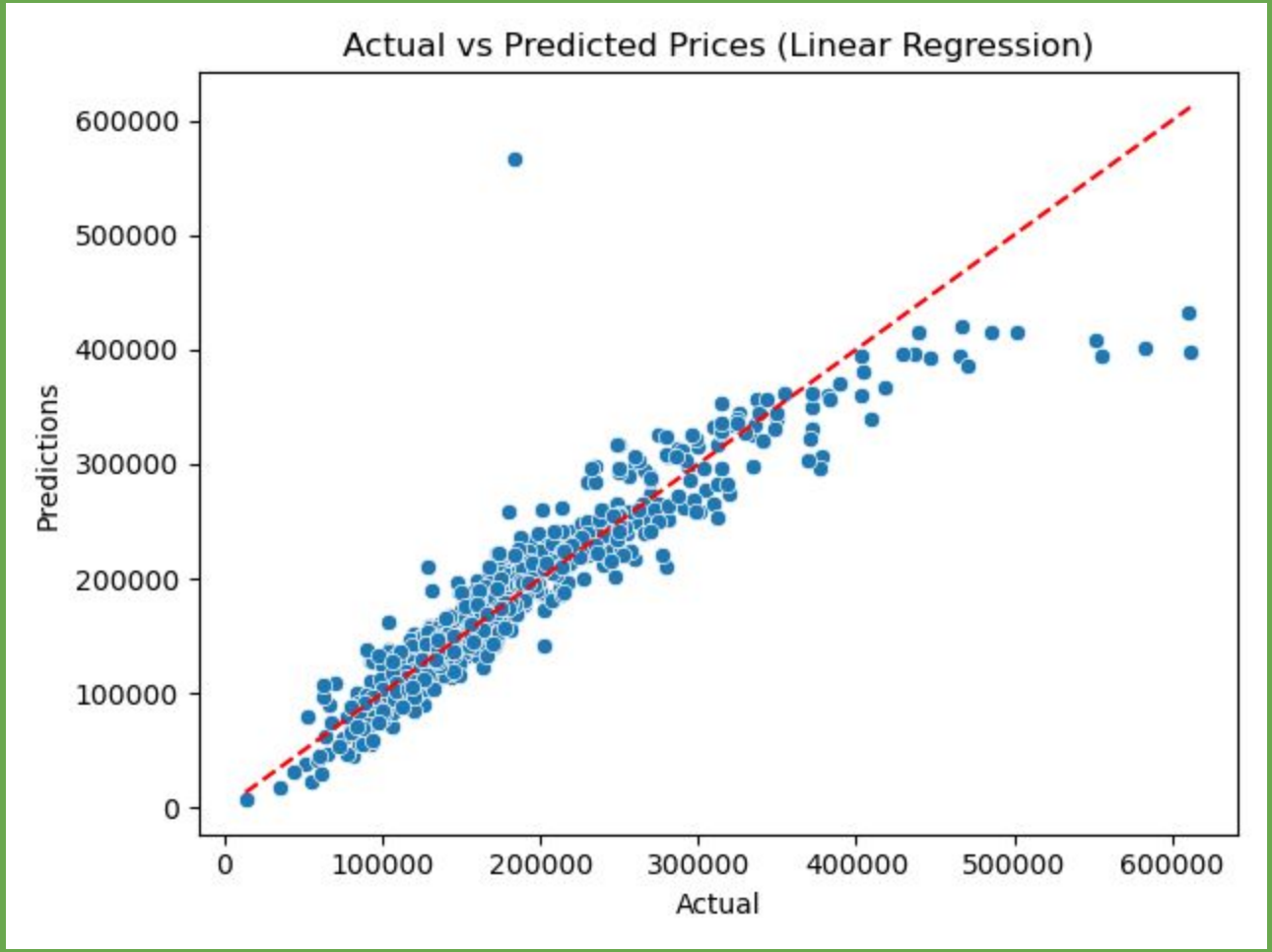
Model Training

- 75/25 split for training and testing sets, as this had the best balance.
- 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 - Linear Regression

- 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.

Recommendations

- Bring in professional services to improve quality of the house.
- Improve neighborhood conditions


QUESTIONS