House Price Prediction Using LinearRegression Model.

By: Ayodeji Ejigbo

Table of content

- Introduction
- Dataset Overview
- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Model Development and Evaluation
- Conclusion
- Future Work

Introduction

This project aims to develop a regression model to predict house prices in Ames, IA, using the Ames housing dataset. The project involves exploring the dataset, preprocessing the data, building several regression models, and selecting the best model for prediction.

Dataset Overview

The dataset used in this project is sourced from Kaggle and consists of two files: train.csv and test.csv.

train.csv

- Number of Variables: 81
- Number of Observations: 2051
- Description: This file contains detailed information about various features of houses in Ames, IA, along with their respective sale prices. Each row represents a unique house, and the columns represent different attributes such as lot size, building type, year built, and more, including the target variable SalePrice.

test.csv

- Number of Variables: 80
- Number of Observations: 878
- Description: This file includes the same features as the training set, except for the SalePrice column, which needs to be predicted. Each row represents a unique house, and the columns provide information on various attributes similar to those in the training set.
-

Baseline Predictions

These are the baseline metrics and their values.

Metrics	Value
baseline_predictions	180779.0657
mean_absolute_error	60143.5905380776
mean_squared_error	6687232614.760764
bp_r^2_score	-0.0011414159496483339

Data Processing

Missing Values (Top 5)

Column Name	Count
Exterior material quality	2051
Central Air Condition	2051
Exterior material quality	1986
Paved driveway	1911
Central air conditioning	1651

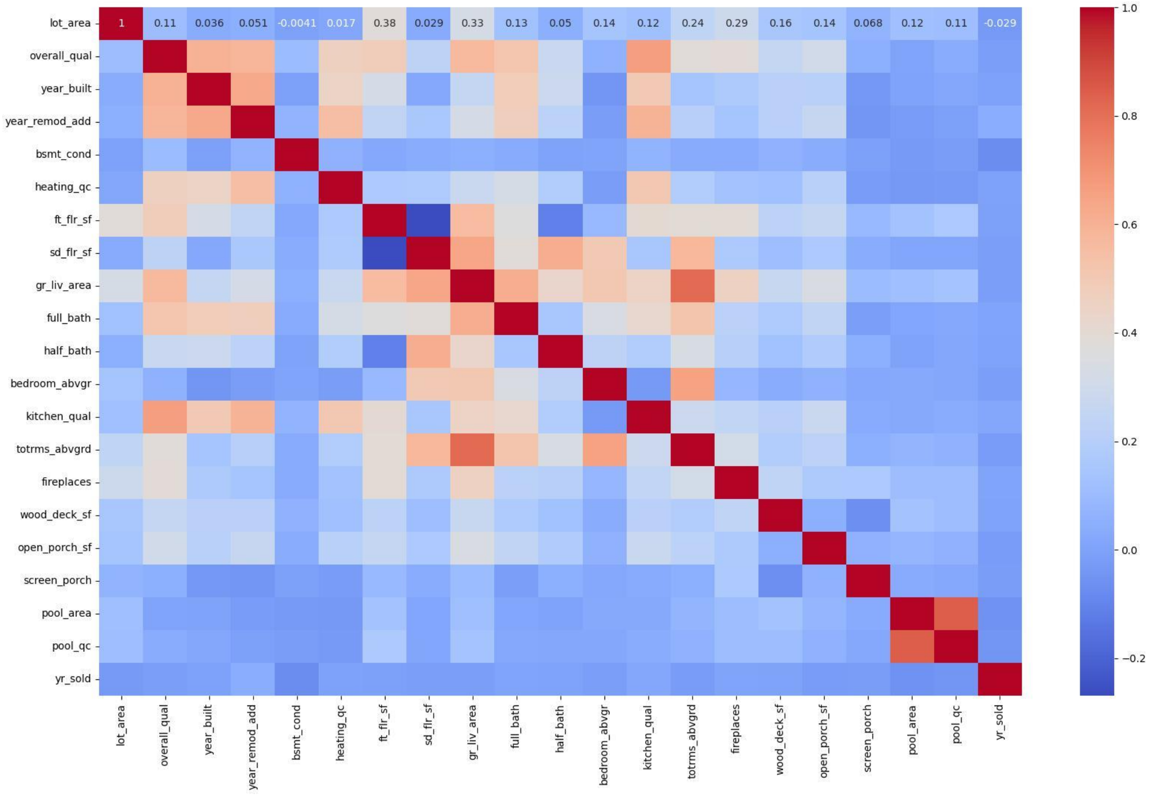
Data Processing

Defining Scales For Some Categorical Columns:

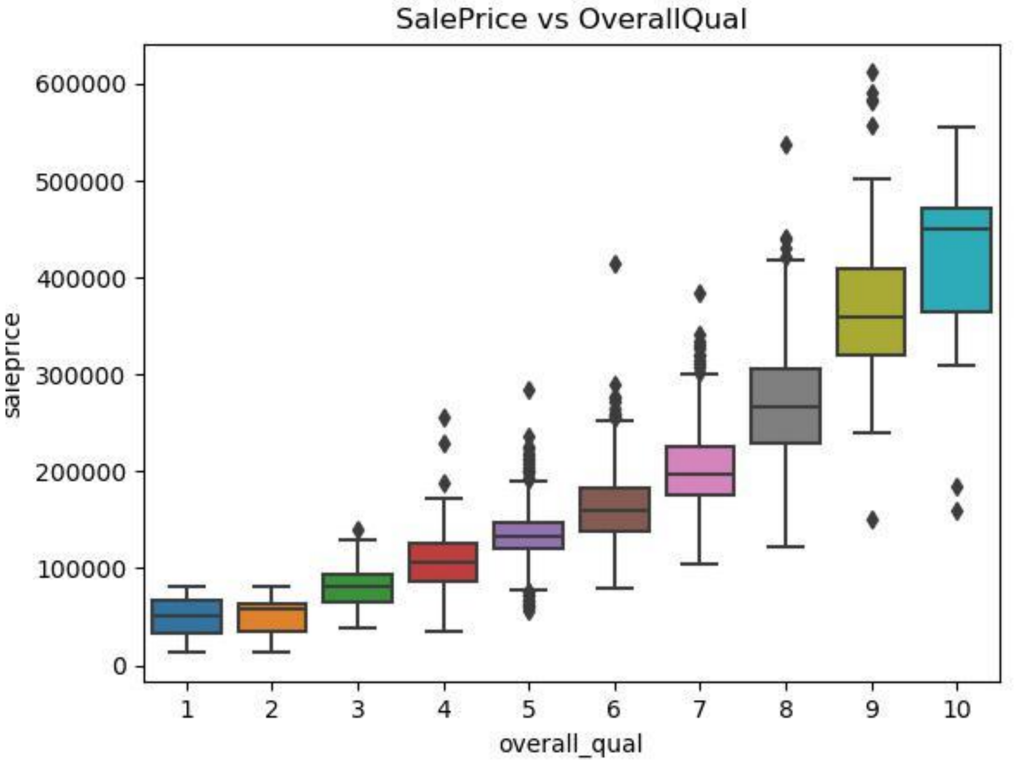
- Fireplace quality
- Height of the basement
- General condition of the basement
- Kitchen quality
- Heating quality and condition
- Garage quality
- Pool quality

Exploratory Data Analysis (EDA)

HeatMap Showing Relationships Between Numerical Variables

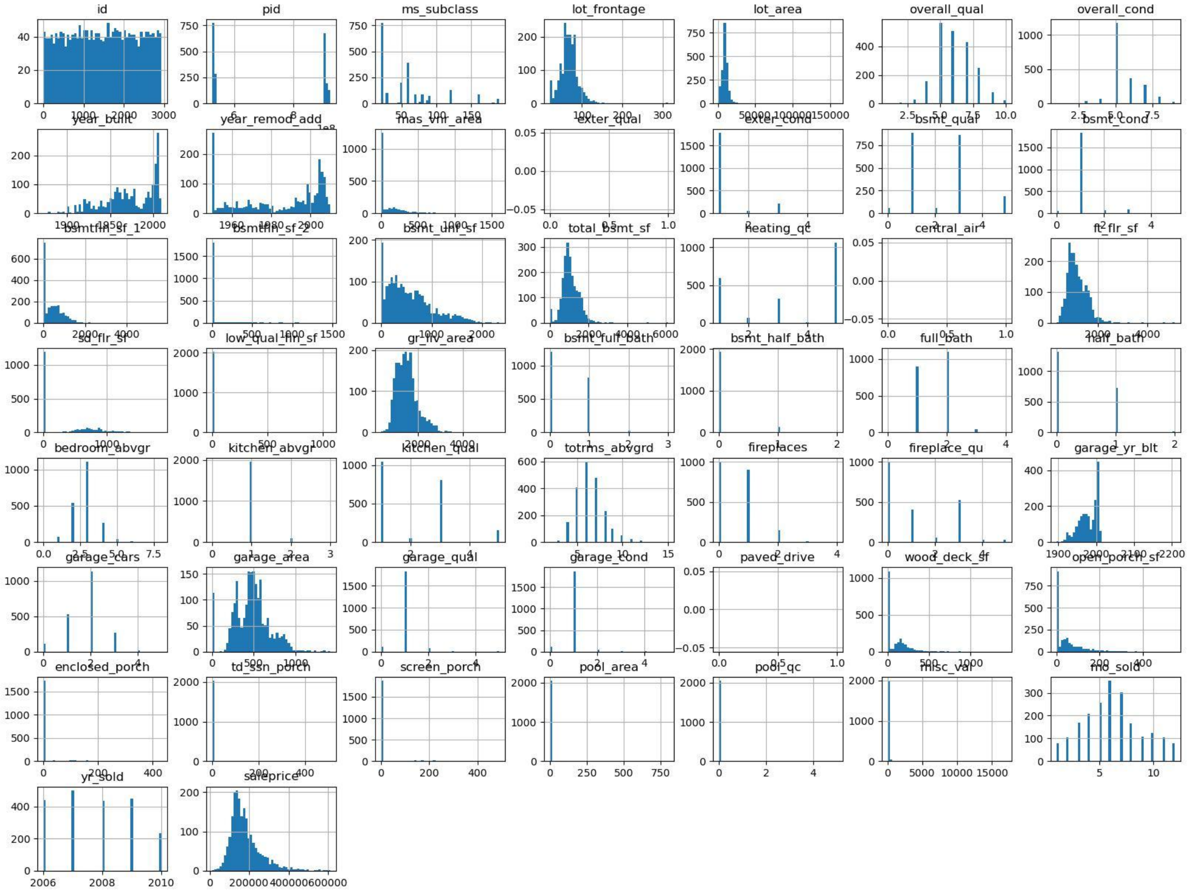


Sales Price VS Overall Quality



This shows the relationship between sales price and the overall Quality.

Numerical Histograms



Preprocessing

- **SimpleImputer()** - To fill in missing values in a dataset using a specified strategy (e.g., mean, median, most frequent)
- **StandardScaler()** - This is to the features by removing the mean and scaling to unit variance, ensuring each feature has a mean of 0 and a standard deviation of 1.
- **OneHotEncoder** - To convert categorical variables into a binary (one-hot) encoded format, creating a new binary column for each category.

Model Development and Evaluation

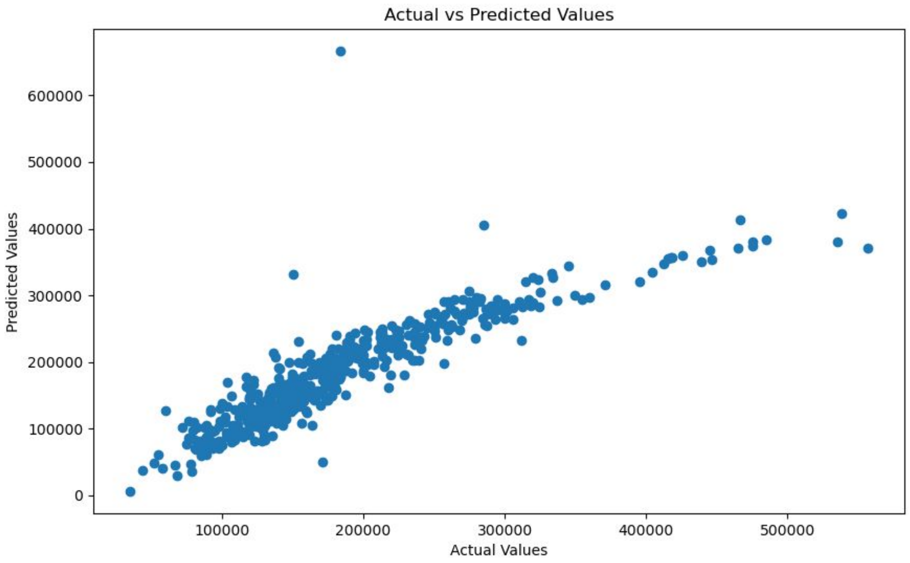
Model: Linear Regression

Metrics	Value
baseline_predictions	23155.53
mean_absolute_error	1442922420.82
mean_squared_error	37985.81
bp_r^2_score	0.783

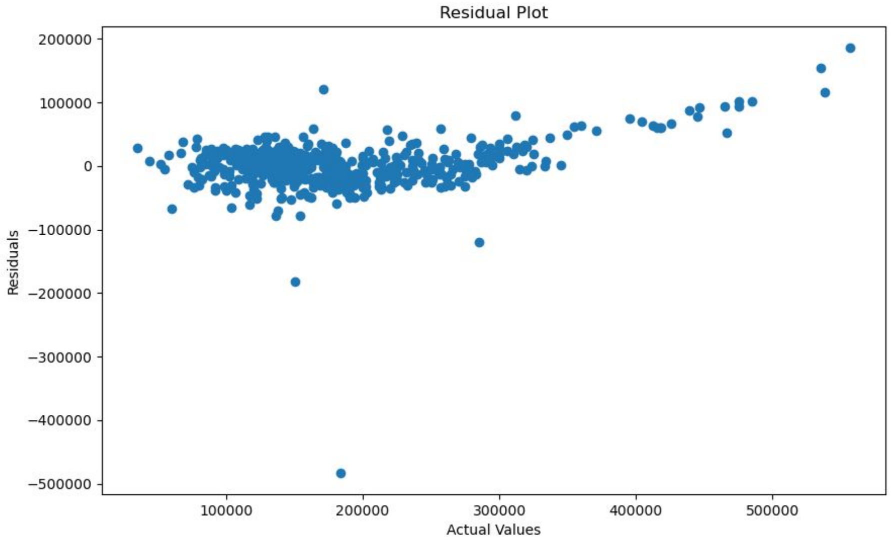
Baseline Model Vs Linear Regression Model

	Baseline Model	Linear Regression Model
mean_absolute_error	180779.06	23155.53
mean_squared_error	60143.5905380776	1442922420.82
root_mean_squared_error	81775.50131158331	37985.81
bp_r^2_score	-0.0011	0.783

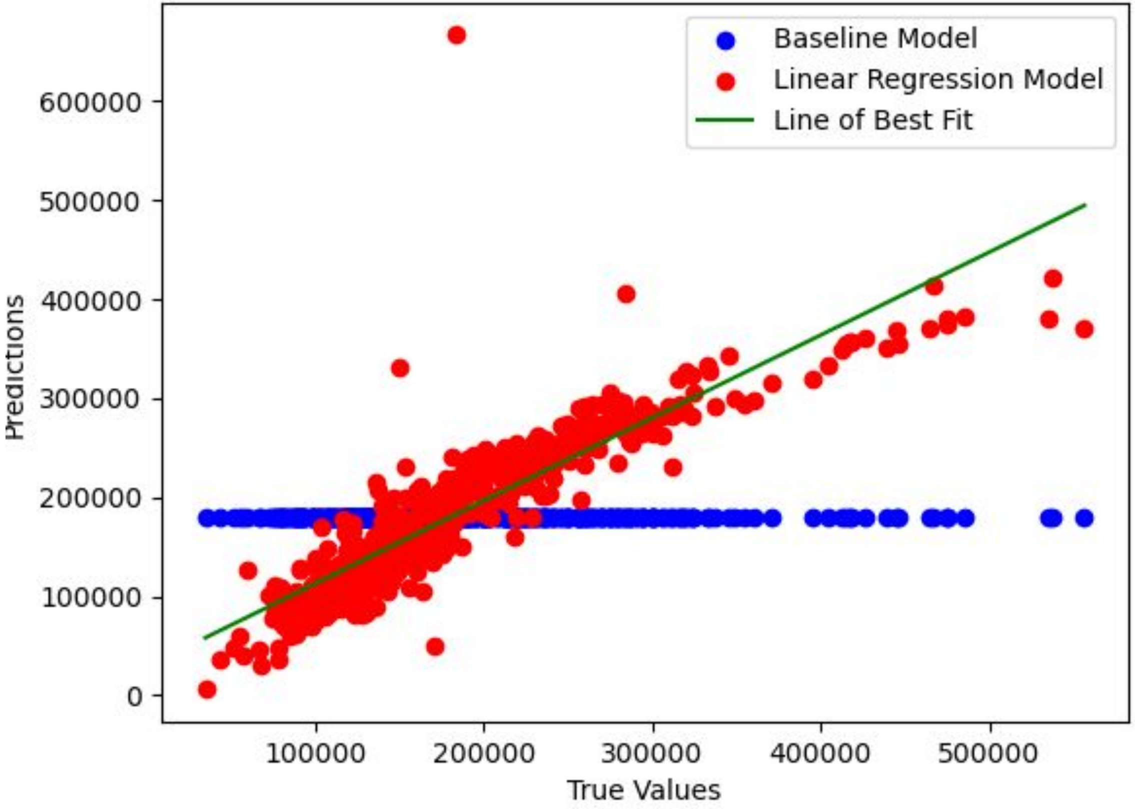
Actual vs Predicted Values



Residual Plot



Baseline Model Vs Linear Regression Model



Conclusion

- My analysis indicates that features like overall quality, above grade (ground) living area square feet, and basement have strong correlations with Sales price.

- Also, the linear regression model outperforms the baseline model, showing predictions that closely align with true values, as evidenced by the scatter plot.

- The data distributions further confirm the significance of these features, with OverallQual showing a clear positive relationship with SalePrice.

- The linear regression model's performance suggests it is a good starting point, but there is room for improvement.

Recommendations

- The main focus should be on the overall quality, above grade (ground) living area square feet, and Basement full bathrooms.
- To improve the model, I will consider advanced regression techniques like Ridge or Lasso regression.

Thank You!