Project 3: Predicting House Prices Using Regression

Introduction

For this project, I aim to build a regression model to predict house prices using the dataset provided in the Kaggle House Prices Competition. The dataset contains various features related to house characteristics, such as square footage, number of rooms, location, and more. The objective is to train a model that accurately estimates house prices based on these attributes.

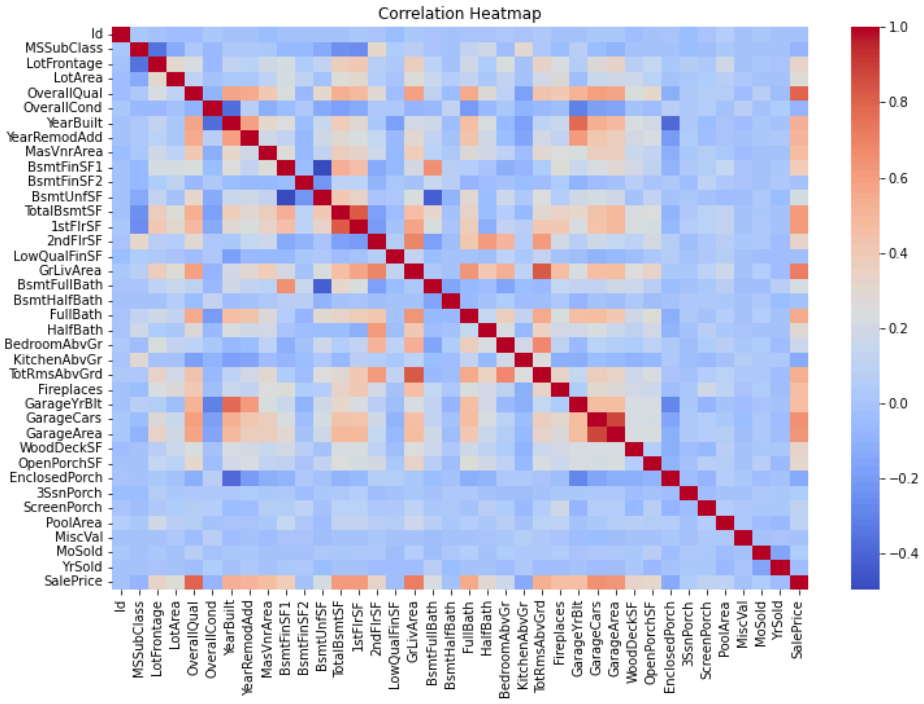
Data Preprocessing

The dataset consists of 4 files The dataset consists of four files: data_description.txt, train.csv, test.csv, and sample_submission.csv. To prepare the data for modeling, I did the following preprocessing steps:

1. Handled duplicates & missing values
2. Feature encoding
3. Feature scaling

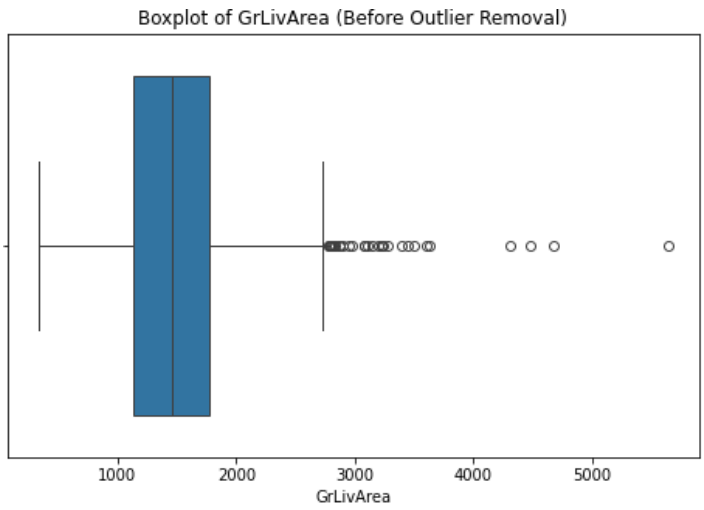
Experiment 1: Linear Regression Model

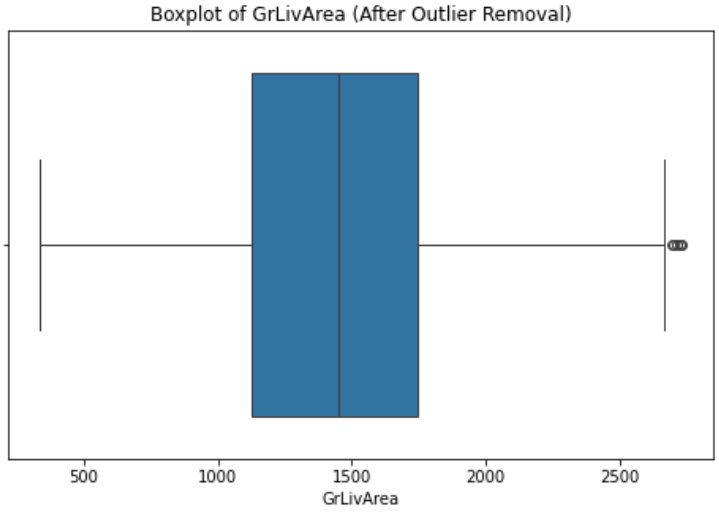
Before modeling, I conducted exploratory data analysis to understand the dataset. I generated a correlation heatmap to visualize relationships between features and the target variable (SalePrice). Highly correlated features such as OverallQual, GrLivArea, and TotalBsmtSF suggest that these variables significantly impact house prices.



	OverallQual	GrLivArea	TotalBsmtSF	SalePrice
OverallQual	1.000000	0.593007	0.537808	0.790982
GrLivArea	0.593007	1.000000	0.454868	0.708624
TotalBsmtSF	0.537808	0.454868	1.000000	0.613581
SalePrice	0.790982	0.708624	0.613581	1.000000

A scatter plot showing the relationship between living area (GrLivArea) and sale price (SalePrice) showed some extreme outliers. I removed these outliers to avoid skewing the model. These outliers were houses that had very large living areas but relatively low prices, suggesting they might be unusual cases.





I built a linear regression model using scikit-learn. I split the dataset into training and validation sets to test the model on new data. After training, I made predictions for both the validation set and the test dataset. Since the test dataset does not have actual SalePrice values, I evaluated the model's performance using Root Mean Squared Error (RMSE) on the validation set. The RMSE I calculated was 20,948.79, which shows the prediction error of the model.

Conclusion

This project focused on using regression modeling to predict house prices. I explored the data to find important factors that affect prices and fixed issues like outliers that could distort the results. The linear regression model produced an RMSE of 20,948.79, indicating a moderate level of accuracy in predictions. To improve future results, I could work on better feature engineering, improve the handling of missing values, and try more complex models like Decision Trees or Random Forests.