Housing Price Predictions

Walter Hickman Jr.
Jamie McGraner
Anand Punwani
Anthony Wilson

Project Overview

This project aims to develop predictive models for estimating housing prices using various regression techniques.

By leveraging a comprehensive dataset, the project seeks to identify key factors that influence housing prices and accurately forecast future prices.

Goals

- Accurate Predictions
- Feature Identification
- Model Comparison
- Optimization
- Visualization

Questions

- Which features most significantly affect housing prices?
- How can we handle missing data and outliers?
- Which regression models provide the best performance for this dataset?
- What are the optimal hyperparameters for each model?
- How can scatter plots, heatmaps and box plots be used to visualize data and model results?
- How do time-based trends affect housing prices?

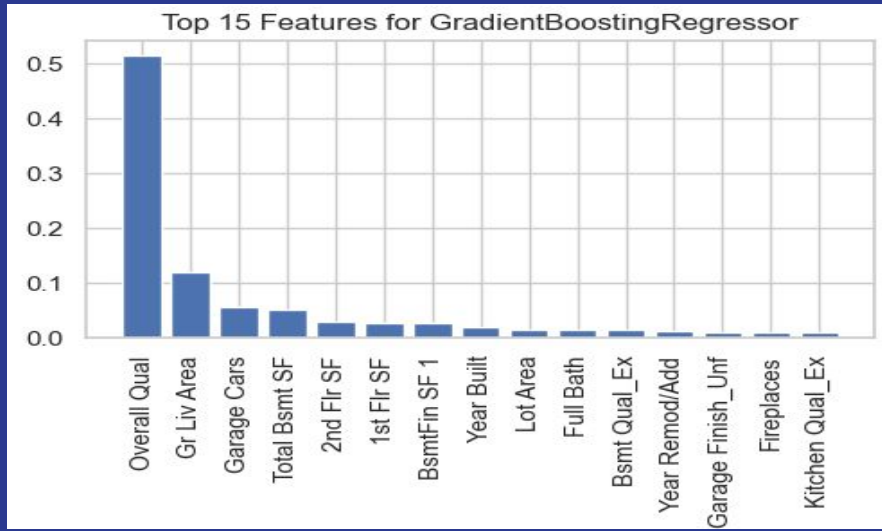
Data Collection & Data Cleaning

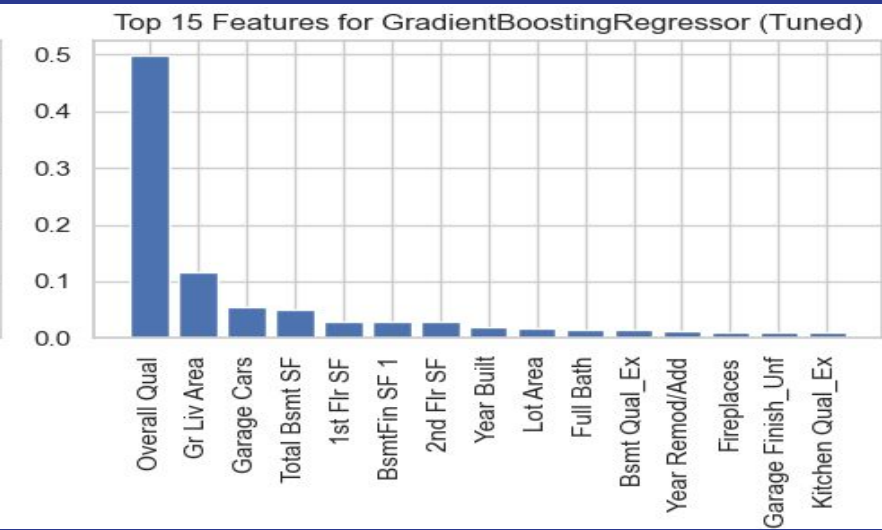
- Data Collection
  - Ames Housing Dataset from Kaggle
- Handling Missing Values
  - Imputation
  - Dropping Columns
- Outlier Detection & Removal
- Label Encoding
- Feature Engineering

Data Exploration

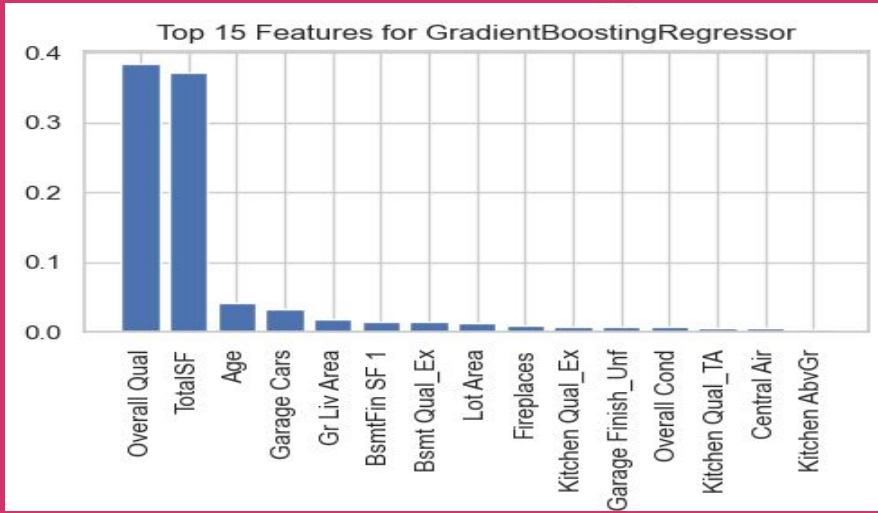
- Descriptive Statistics
- Correlation Analysis
- Data Visualization
  - Histograms
  - Box Plots
  - Scatter Plots
  - Heatmaps
- Target Variable Analysis
  - `SalePrice`
- Feature Relationships

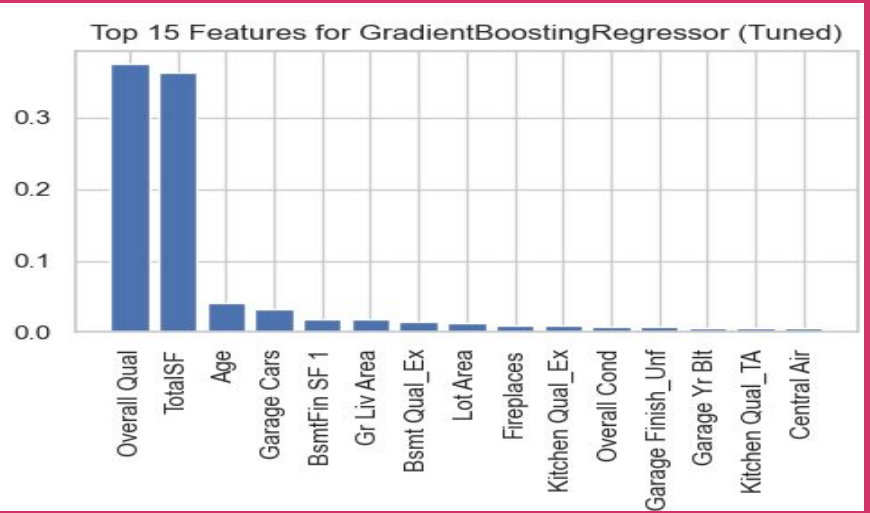
Top 15 Features - Pre Feature Engineering





Top 15 Features - Post Feature Engineering





Summary Statistics

	TotalSF	Age	Overall Qual	Garage Area	Gr Liv Area	Lot Area	SalePrice
count	2930.000000	2930.000000	2930.000000	2930.000000	2930.000000	2930.000000	2930.000000
mean	2546.832707	36.434130	6.094881	472.819734	1499.690444	10147.921843	180796.060068
std	803.772450	30.291357	1.411026	215.009836	505.508887	7880.017759	79886.692357
min	334.000000	-1.000000	1.000000	0.000000	334.000000	1300.000000	12789.000000
25%	2000.500000	7.000000	5.000000	320.000000	1126.000000	7440.250000	129500.000000
50%	2452.000000	34.000000	6.000000	480.000000	1442.000000	9436.500000	160000.000000
75%	2990.000000	54.000000	7.000000	576.000000	1742.750000	11555.250000	213500.000000
max	11752.000000	136.000000	10.000000	1488.000000	5642.000000	215245.000000	755000.000000

Goal 1: Accurately Predict Housing Prices

To achieve this goal, we followed a systematic approach involving the following steps:

- Data Collection
- Data Cleaning & Preprocessing
- Feature Engineering
- Exploratory Data Analysis
- Model Building
- Model Evaluation
- Model Deployment

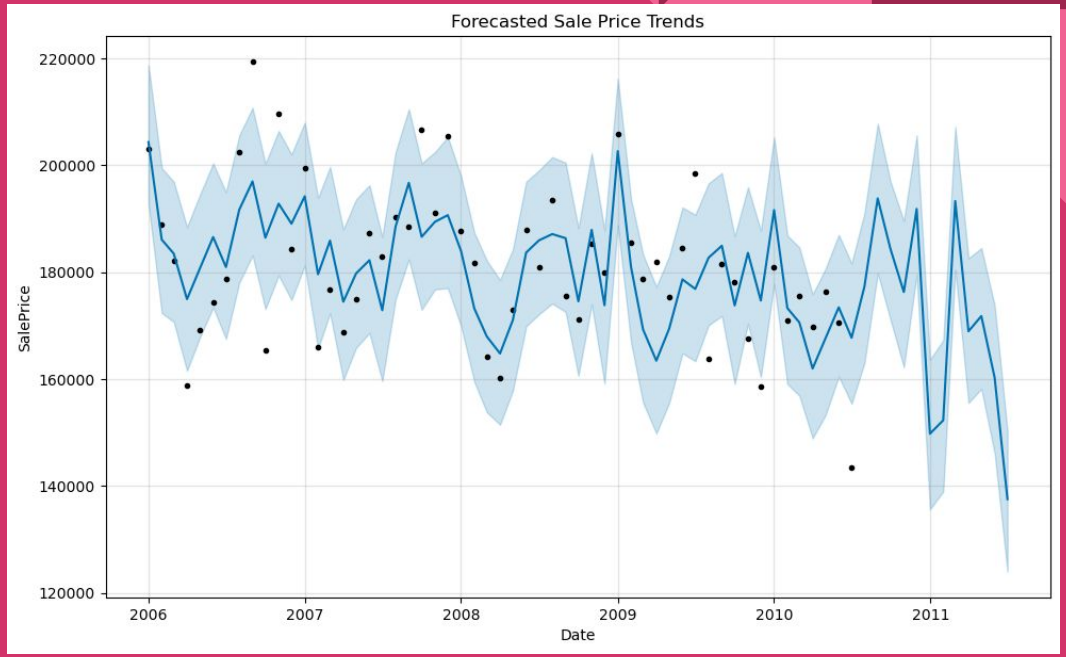
Conclusion 1: Accurate Housing Price Predictions

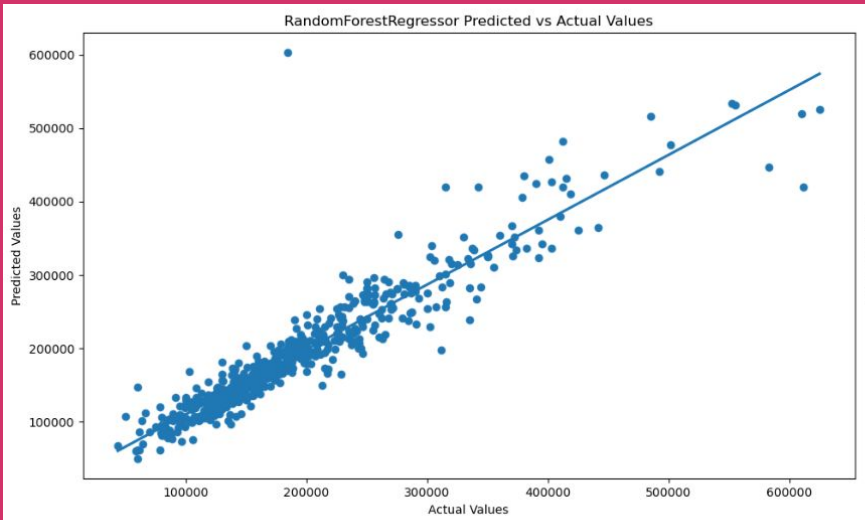
	Model	Train Score	Test Score	MSE	MAE	R2	Top Features	Predicted Values
0	LinearRegression	0.789	0.803	1583096593.894	25002.181	0.803	Not applicable for this model	[187536.00612379232, 128051.58825578082, 22081...
1	LinearRegression (Tuned)	0.789	0.803	1583096593.894	25002.181	0.803	Not applicable for this model	[187536.00612379232, 128051.58825578082, 22081...
2	DecisionTreeRegressor	1.000	0.757	1950634408.889	26676.053	0.757	[(0.5239780146313097, Overall Qual), (0.304986...	[182000.0, 108000.0, 196500.0, 132000.0, 11900...
3	DecisionTreeRegressor (Tuned)	0.896	0.807	1547925389.544	22935.049	0.807	[(0.5817153022616219, Overall Qual), (0.318472...	[183021.73913043478, 105657.41414141415, 20258...
4	RandomForestRegressor	0.979	0.879	969688155.213	18128.593	0.879	[(0.49338989121568316, Overall Qual), (0.34226...	[175460.0, 104664.66333333332, 198032.04, 1228...
5	RandomForestRegressor (Tuned)	0.981	0.875	1004591187.282	18246.411	0.875	[(0.5098250000444117, Overall Qual), (0.326676...	[184664.23466666666, 103516.62698412697, 19912...
6	GradientBoostingRegressor	0.929	0.890	880869704.963	18320.882	0.890	[(0.4424683916223565, Overall Qual), (0.406988...	[180038.32805481498, 106649.90554448347, 20288...
7	GradientBoostingRegressor (Tuned)	0.929	0.890	880869704.963	18320.882	0.890	[(0.4424683916223565, Overall Qual), (0.406988...	[180038.32805481498, 106649.90554448347, 20288...

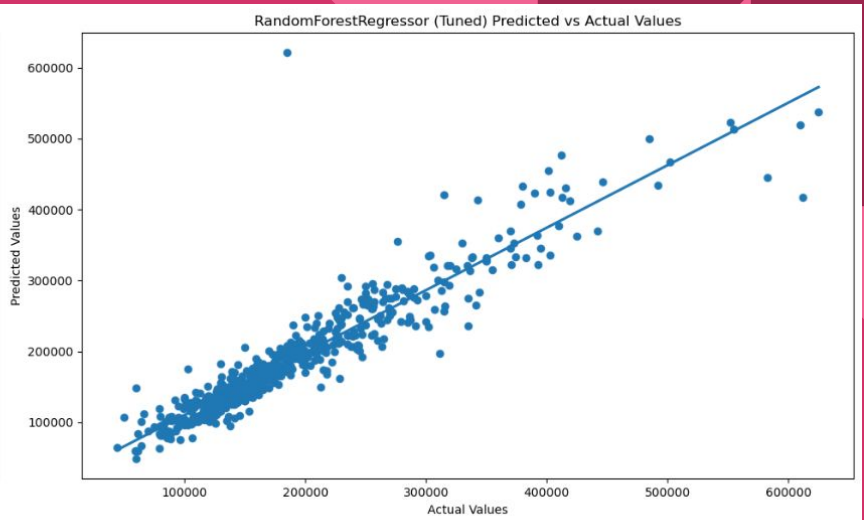
Conclusion 1

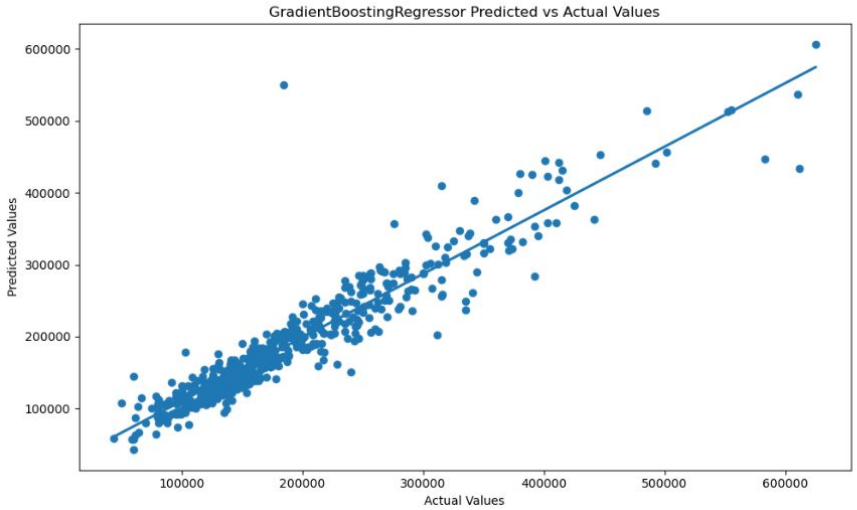
Accurate Housing Price Predictions:

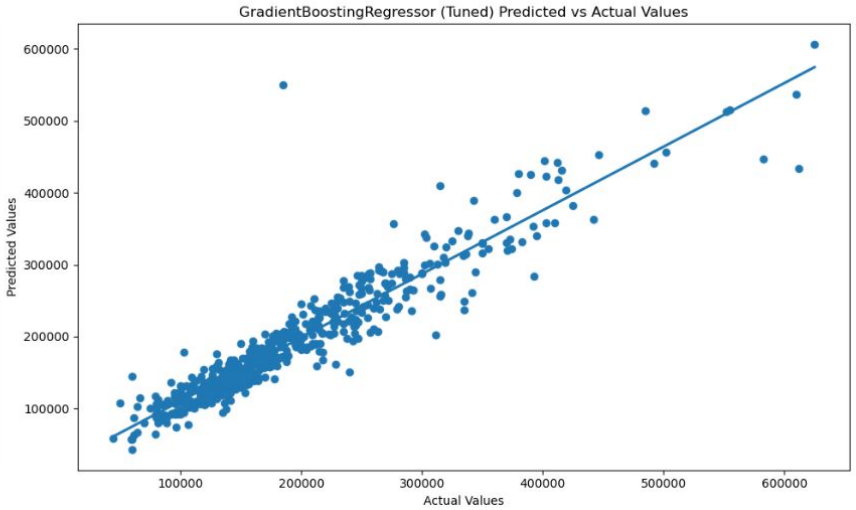
- The project successfully developed predictive models with a high degree of accuracy.




RandomForestRegressor Predicted vs Actual Values
Predicted Values
Actual Values
100000
200000
300000
400000
500000
600000


RandomForestRegressor (Tuned) Predicted vs Actual Values
Predicted Values
Actual Values
100000
200000
300000
400000
500000
600000

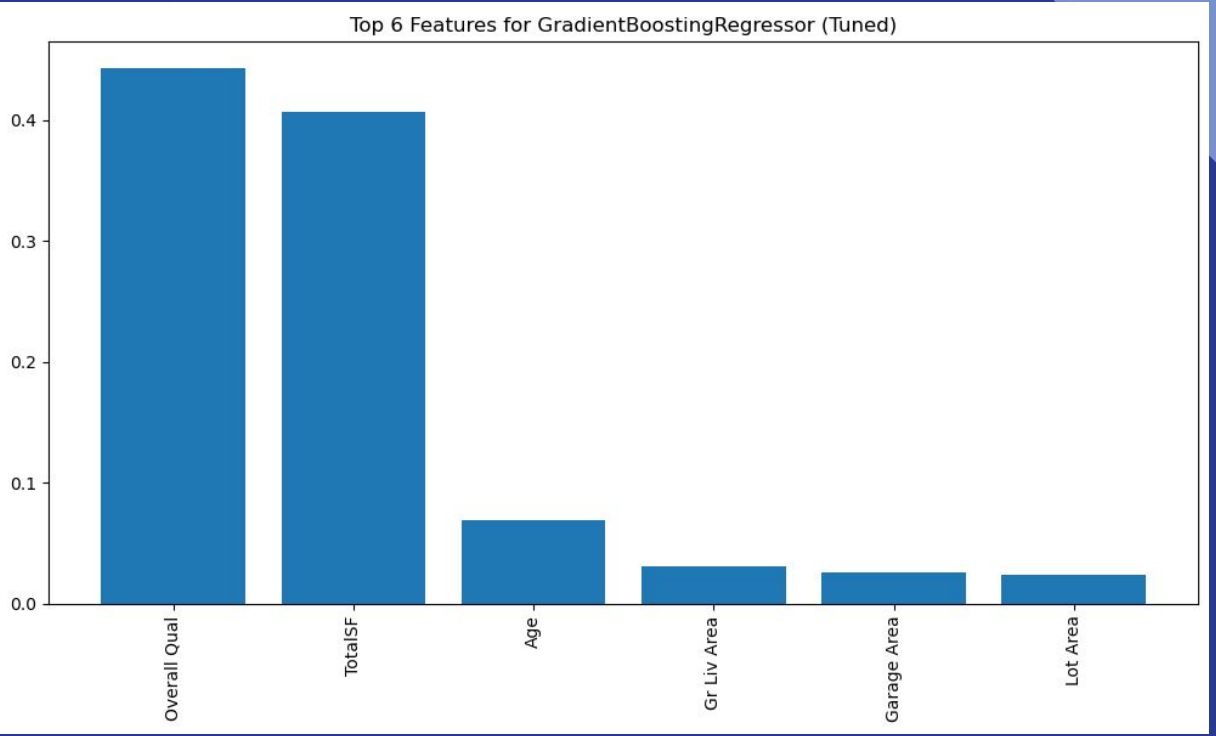

GradientBoostingRegressor Predicted vs Actual Values
Predicted Values
Actual Values
100000
200000
300000
400000
500000
600000


GradientBoostingRegressor (Tuned) Predicted vs Actual Values
Predicted Values
Actual Values
100000
200000
300000
400000
500000
600000

Goal 2: Identify Key Features Affecting Housing Prices

We took the following steps:

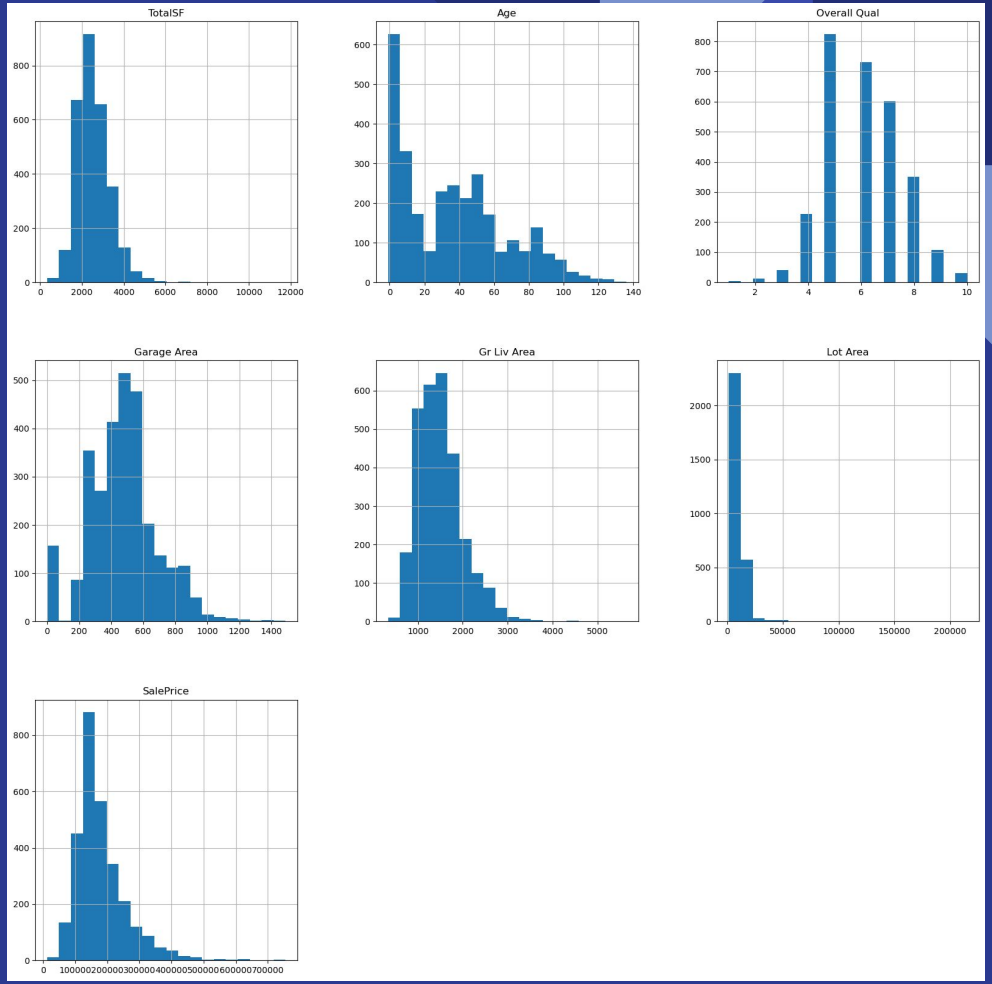
- Correlation Analysis
- Feature Importance
- Visual Analysis
- Interpretability



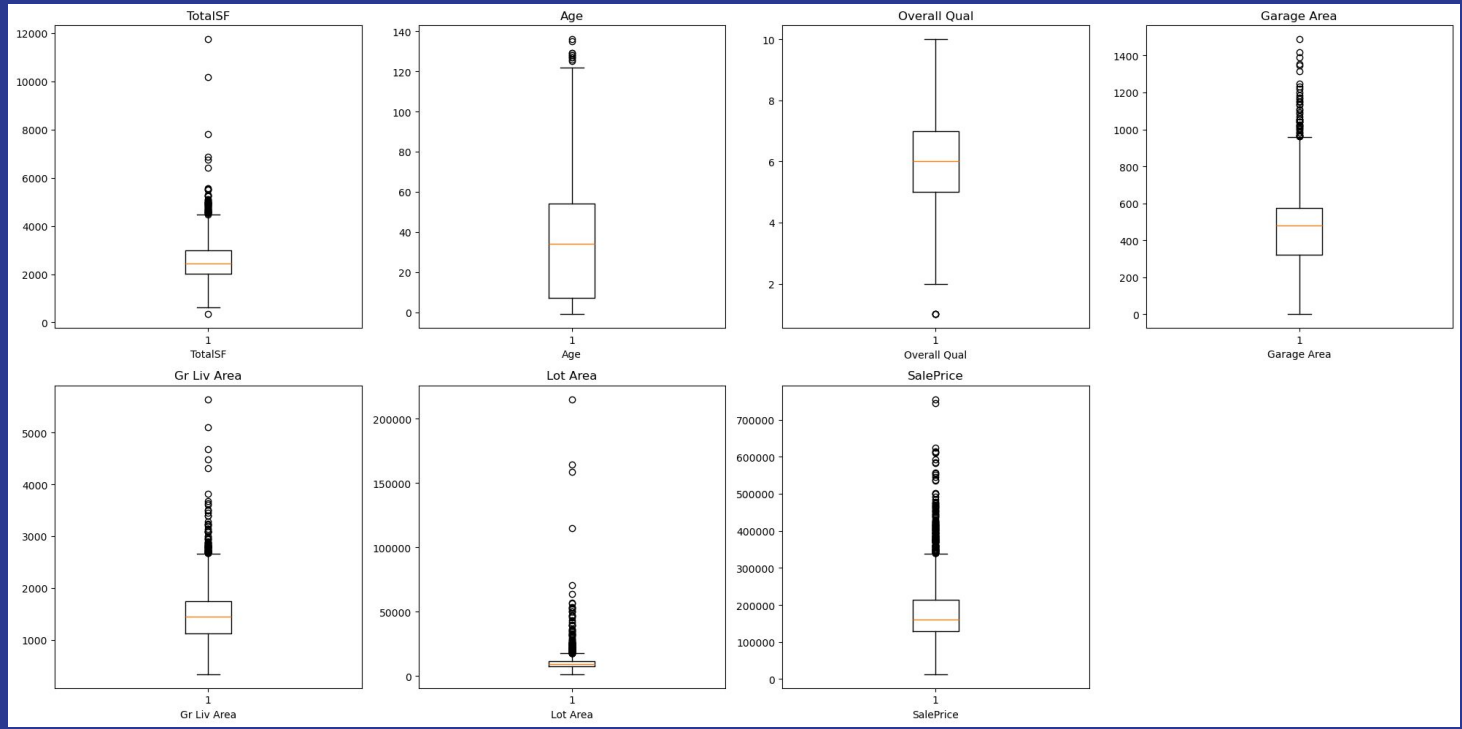
Conclusion 2

Identification of Key Features Impacting Housing Prices:

- TotalSF
  - First Floor SF
  - Second Floor SF
  - Total Basement SF
- Age
  - Year Sold - Year Built
- Overall Qual
  - 1-10 Scale
- Garage Area
  - Square Footage
- Gr Liv Area
  - Square Footage
- Lot Area
  - Square Footage
- SalePrice



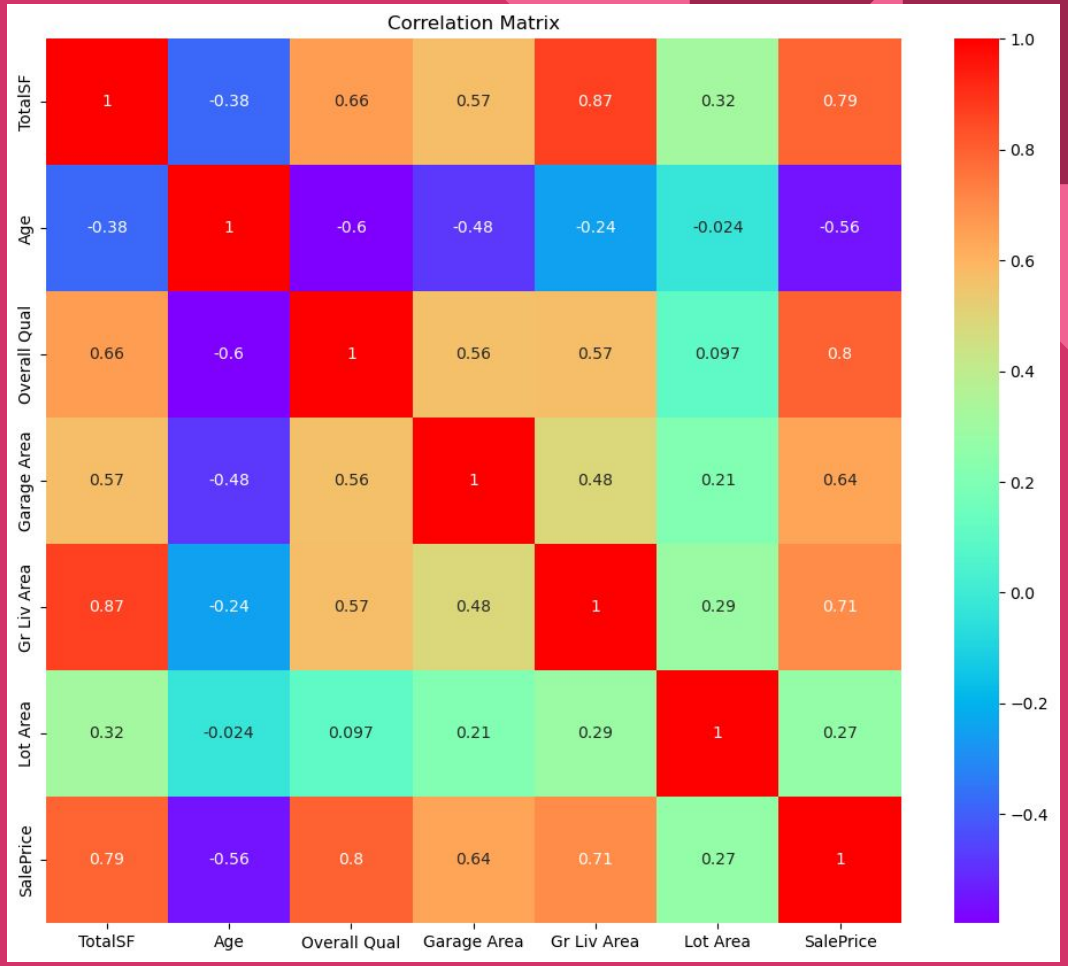
Conclusion 2: Identification of Key Features Impacting Housing Prices



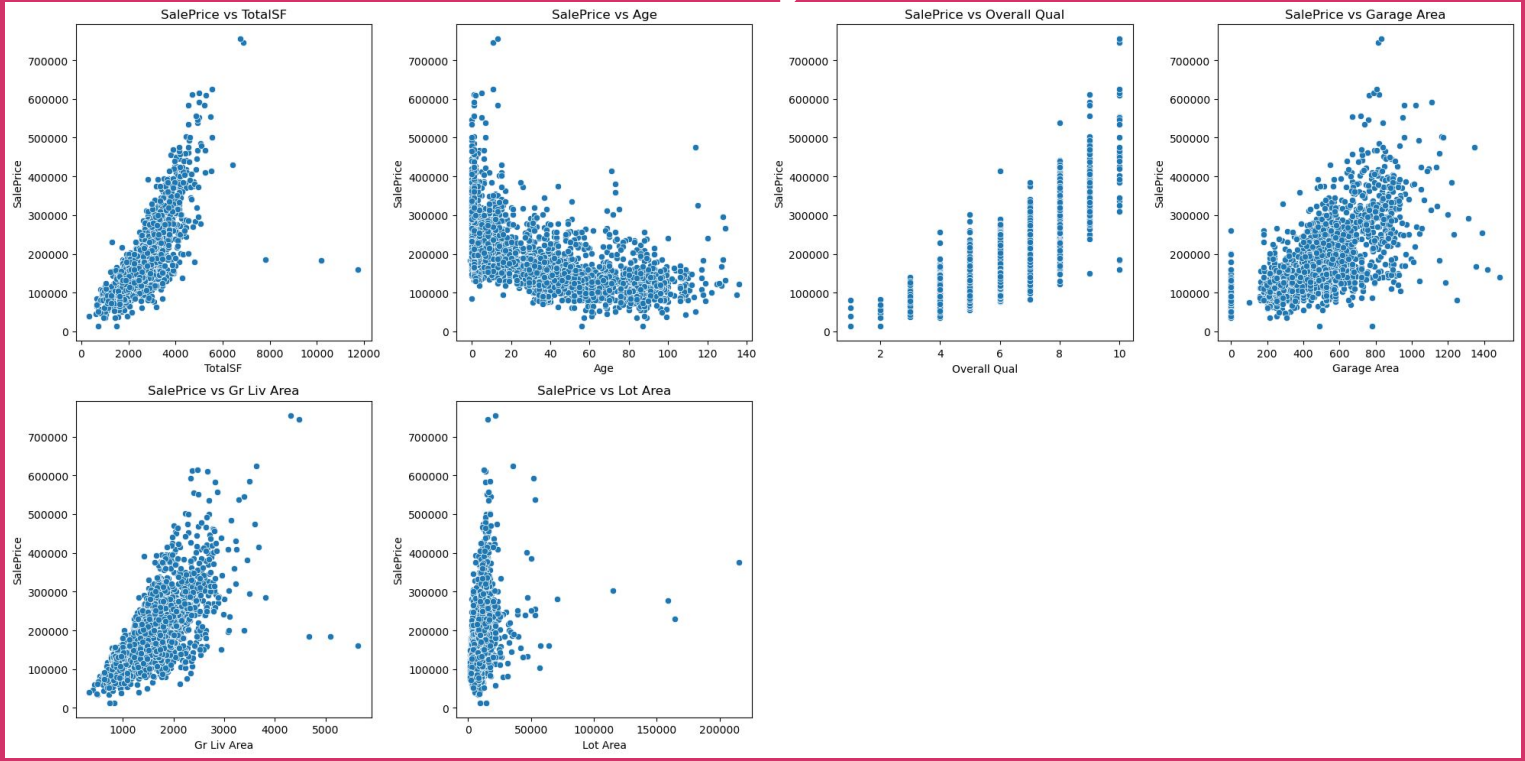
Conclusion 3

Insights into Housing Market Dynamics:

- Newer homes in well-developed neighborhoods with larger living spaces command higher prices.
- Features such as proximity to amenities and quality of construction were found to enhance property value.



Conclusion 3: Insights into Housing Market Dynamics



Summary

- Our project focuses on predicting housing prices using advanced machine learning models.
- By analyzing a comprehensive dataset from Kaggle, we developed multiple regression models.
- This project not only provides accurate price predictions but also offers valuable insights into the factors affecting housing prices, making it a practical tool for real estate professionals and potential home buyers.

Problems Encountered

During the project, we encountered several challenges that impacted our workflow such as:

- Finding a suitable API that could effectively complement our dataset
- Integrating multiple datasets due to inconsistencies in column structures and missing data.

Future Considerations

- Finding a more comprehensive dataset(s) with more up to date data
- Testing future data with the Prophet model to see if its predictions were correct
- Finding more datasets that work well together when combined

The End