PREDICTING HOUSE PRICES USING THE HOUSE PRICES-ADVANCED REGRESSION TECHNIQUES DATASET

Ene-Obong Daniella
(13/08/2024)

TABLE OF CONTENTS

I. DEFINITION

A. Project Overview

Unlike with many consumer goods which have a short lifespan, the owners. stakeholders and key players of real estate generally realise their benefits over a long period of time. Therefore, accurately valuing property requires taking into consideration the economic, social, governmental and environmental factors that altogether influence the set prices of real estate.[1] Many features of the real estate are taken into account in this valuation. For instance, in the valuation of houses for sale, features like the lot size, accessible garages, rooms available and so on are analysed before a definitive price is set.

The primary objective of this project is to develop a robust predictive model capable of accurately estimating house prices based on a comprehensive dataset encompassing various property attributes.

The project involves a thorough exploration of the dataset, including data cleaning, feature engineering, and the application of appropriate machine learning algorithms. The model's performance will be evaluated using relevant metrics, and insights derived from the analysis will be presented.

This report is intended for data scientists, real estate professionals, and individuals interested in understanding the factors influencing house prices.

B. Related Work

As a wildly interesting and rewarding sector with sufficient recorded data, predicting house prices has been a subject of interest for both academic researchers and industry practitioners. Numerous studies have explored various machine learning techniques and feature engineering approaches to tackle this problem.

Some studies investigate this subject using linear regression, Random Forest Regressor, Catboost, XGBoost, etc.[2] [3] [4] [5]

C. Problem Statement

When preparing a property for sale, determining the optimal pricing is a crucial step for the homeowner or the real estate agent entrusted with the sale. Expert intuition as the conventional comparable sale analysis can be subjective and prone to errors. This can lead to inaccurate pricing which could in turn lead to a huge loss in profits and commission for both the homeowner and the agent due to undervaluation as well as a prolonged time spent on the market due to overestimation.

The problem to be solved, ***predicting house prices based on various features***, is a regression task using supervised learning for a data-driven approach to property value estimation.

D. Evaluation Metrics

A number of regression metrics are necessary to quantify the performance of the benchmark and solution models on the data.[6] In regression models, the accuracy metric is not as important as regression metrics which provide a genuine and approximate measurement on how well the regression model is able to predict continuous outcomes.

To assess the performance of the developed models, the Mean Absolute Error (MAE), which calculates the average absolute difference between predicted and actual house prices, Root Mean Squared Error (RMSE), the square root of MSE, providing a more interpretable error metric in the original units of the target variable, and R-squared, which calculates the average absolute difference between predicted and actual house prices, would be employed.

II. ANALYSIS

A. Data Exploration

The data used in this project was sourced from the Kaggle competition, House Prices - Advanced Regression Techniques [7]. The dataset used for this project contains 1460 observations of 81 columns of features, providing a comprehensive overview of various house attributes. Initial exploration revealed the presence of a mix of numerical and categorical variables, missing values in certain columns, requiring appropriate imputation techniques, potential outliers in numerical features through statistical analysis and visualisation and possible feature correlations.

Feature	Description	Data Type
MSSubClass	Identifies the type of dwelling involved in the sale.	discrete, numeric
MSZoning	Identifies the general zoning classification of the sale.	categorical
LotFrontage	Linear feet of street connected to property	continuous, numeric
LotArea	Lot size in square feet	continuous, numeric
Street	Type of road access to property	categorical
Alley	Type of alley access to property	categorical
LotShape	General shape of property	categorical
LandContour	Flatness of the property	categorical
Utilities	Type of utilities available	categorical
LotConfig	Lot configuration	categorical
LandSlope	Slope of property	categorical
Neighborhood	Physical locations within Ames city limits	categorical
Condition1	Proximity to various conditions	categorical
Condition2	Proximity to various conditions (if more than	categorical

	one is present)	
BldgType	Type of dwelling	categorical
HouseStyle	Style of dwelling	categorical
1.5Fin One and one-half story	2nd level finished	unknown
1.5Unf One and one-half story	2nd level unfinished	unknown
2.5Fin Two and one-half story	2nd level finished	unknown
2.5Unf Two and one-half story	2nd level unfinished	unknown
OverallQual	Rates the overall material and finish of the house	discrete, numeric
OverallCond	Rates the overall condition of the house	discrete, numeric
YearBuilt	Original construction date	year, date, numeric
YearRemodAdd	Remodel date (same as construction date if no remodeling or additions)	year, date, numeric
RoofStyle	Type of roof	categorical
RoofMatl	Roof material	categorical
Exterior1st	Exterior covering on house	categorical
Exterior2nd	Exterior covering on house (if more than one material)	categorical
MasVnrType	Masonry veneer type	categorical
MasVnrArea	Masonry veneer area in square feet	continuous, numeric
ExterQual	Evaluates the quality of the material on the exterior	categorical
ExterCond	Evaluates the present condition of the material on the exterior	categorical
Foundation	Type of foundation	categorical
BsmtQual	Evaluates the height of the basement	categorical
BsmtCond	Evaluates the general condition of the basement	categorical
BsmtExposure	Refers to walkout or garden level walls	categorical
BsmtFinType1	Rating of basement finished area	categorical
BsmtFinSF1	Type 1 finished square feet	continuous, numeric
BsmtFinType2	Rating of basement finished area (if multiple types)	categorical
BsmtFinSF2	Type 2 finished square feet	continuous,

		numeric
BsmtUnfSF	Unfinished square feet of basement area	continuous, numeric
TotalBsmtSF	Total square feet of basement area	continuous, numeric
Heating	Type of heating	categorical
HeatingQC	Heating quality and condition	categorical
CentralAir	Central air conditioning	categorical
Electrical	Electrical system	categorical
1stFlrSF	First Floor square feet	continuous, numeric
2ndFlrSF	Second floor square feet	continuous, numeric
LowQualFinSF	Low quality finished square feet (all floors)	discrete, numeric
GrLivArea	Above grade (ground) living area square feet	continuous, numeric
BsmtFullBath	Basement full bathrooms	discrete, numeric
BsmtHalfBath	Basement half bathrooms	discrete, numeric
FullBath	Full bathrooms above grade	discrete, numeric
HalfBath	Half baths above grade	discrete, numeric
Bedroom	Bedrooms above grade (does NOT include basement bedrooms)	unknown
Kitchen	Kitchens above grade	unknown
KitchenQual	Kitchen quality	categorical
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)	discrete, numeric
Functional	Home functionality (Assume typical unless deductions are warranted)	categorical
Fireplaces	Number of fireplaces	discrete, numeric
FireplaceQu	Fireplace quality	categorical
GarageType	Garage location	categorical
GarageYrBlt	Year garage was built	year, date, numeric
GarageFinish	Interior finish of the garage	categorical
GarageCars	Size of garage in car capacity	discrete, numeric
GarageArea	Size of garage in square feet	continuous, numeric
GarageQual	Garage quality	categorical
GarageCond	Garage condition	categorical
PavedDrive	Paved driveway	categorical

WoodDeckSF	Wood deck area in square feet	continuous, numeric
OpenPorchSF	Open porch area in square feet	continuous, numeric
EnclosedPorch	Enclosed porch area in square feet	continuous, numeric
3SsnPorch	Three season porch area in square feet	discrete, numeric
ScreenPorch	Screen porch area in square feet	continuous, numeric
PoolArea	Pool area in square feet	discrete, numeric
PoolQC	Pool quality	categorical
Fence	Fence quality	categorical
MiscFeature	Miscellaneous feature not covered in other categories	categorical
MiscVal	$Value of miscellaneous feature	discrete, numeric
MoSold	Month Sold (MM)	date, numeric
YrSold	Year Sold (YYYY)	year, date, numeric
SaleType	Type of sale	categorical
SaleCondition	Condition of sale	categorical
TotalSF	Total square feet	continuous, numeric
TotalBath	Total number of baths	discrete, numeric
HouseAge	Age of the house	year, date, numeric
GaragePresent	Presence of a garage	categorical
SalePrice	Target Variable	continuous, numeric

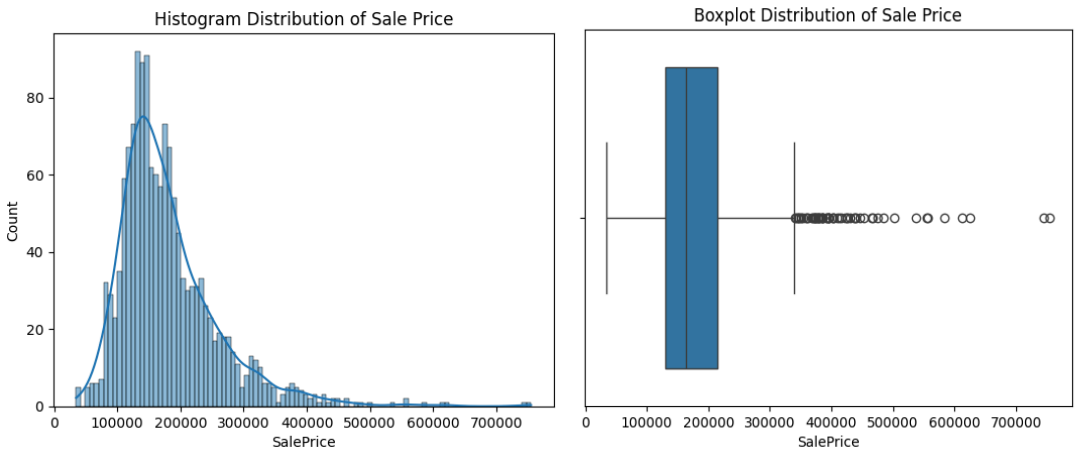
By understanding these fundamental aspects of the data, a solid foundation is laid for subsequent data preprocessing and feature engineering steps.

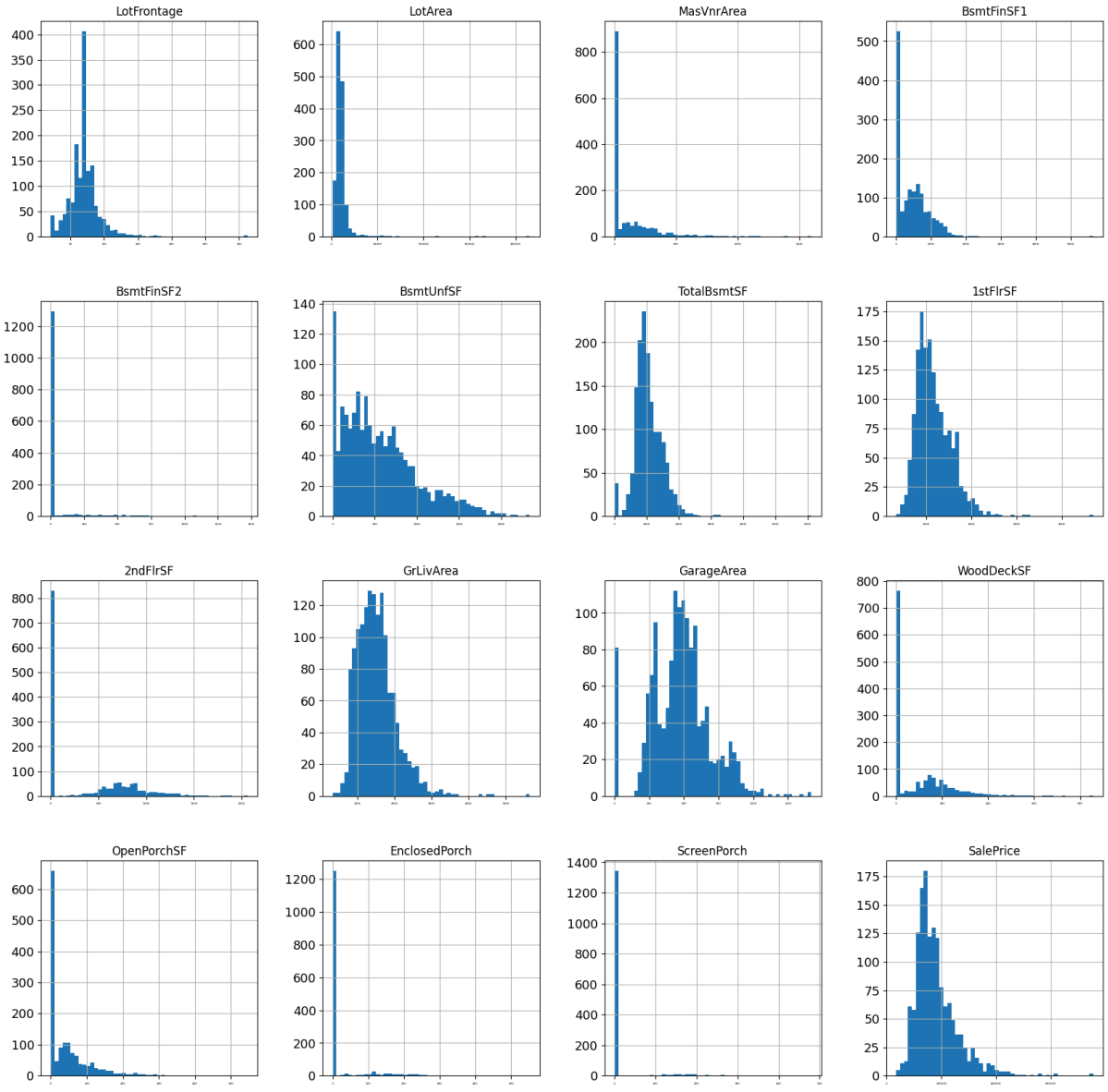
B. Exploratory Visualisation

To gain deeper insights into the data and identify potential patterns, a variety of exploratory visualisations were employed.

1. Univariate analysis

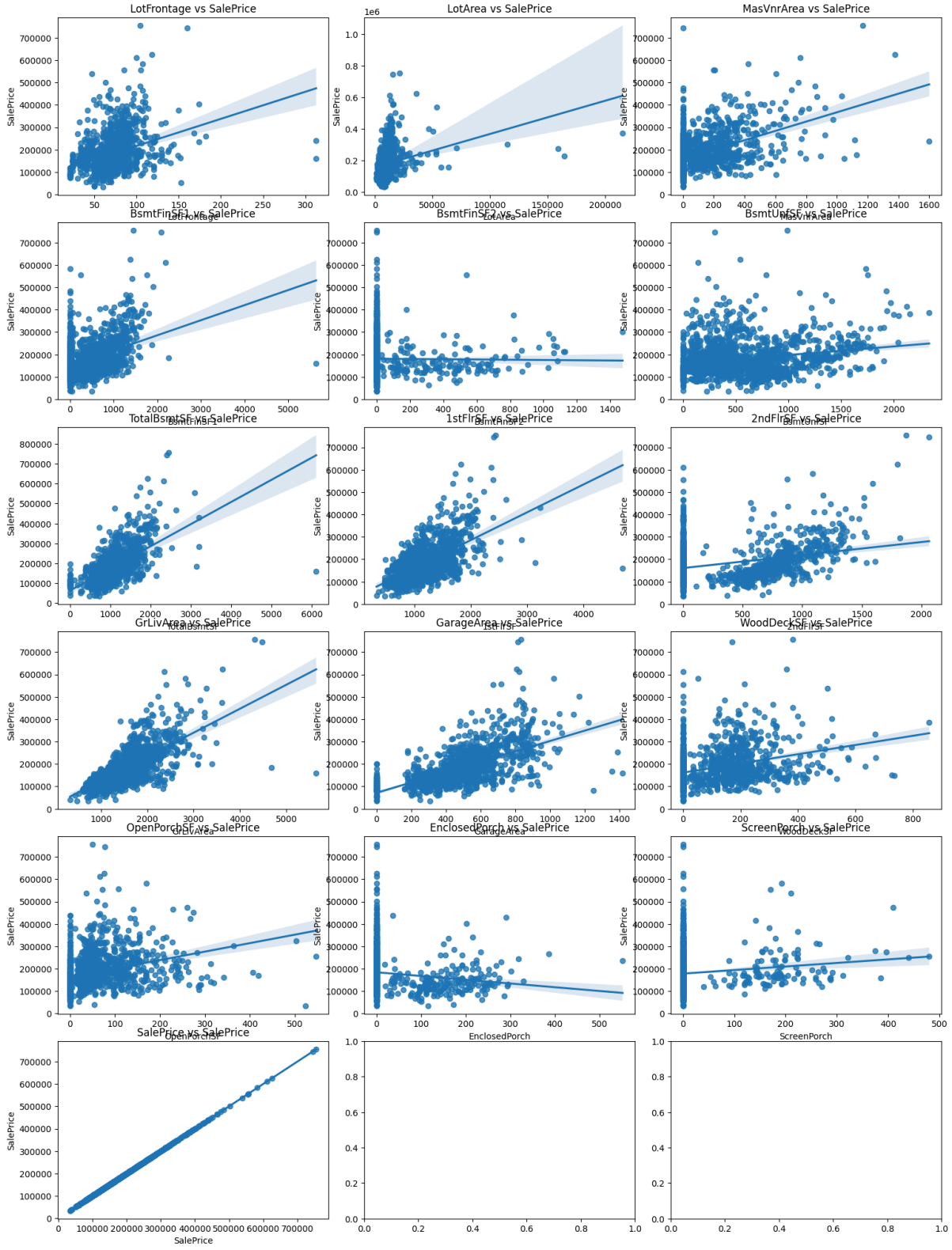
Histograms, box plots, and density plots were used to examine the distribution of numerical features, revealing skewness, outliers, and potential transformations.

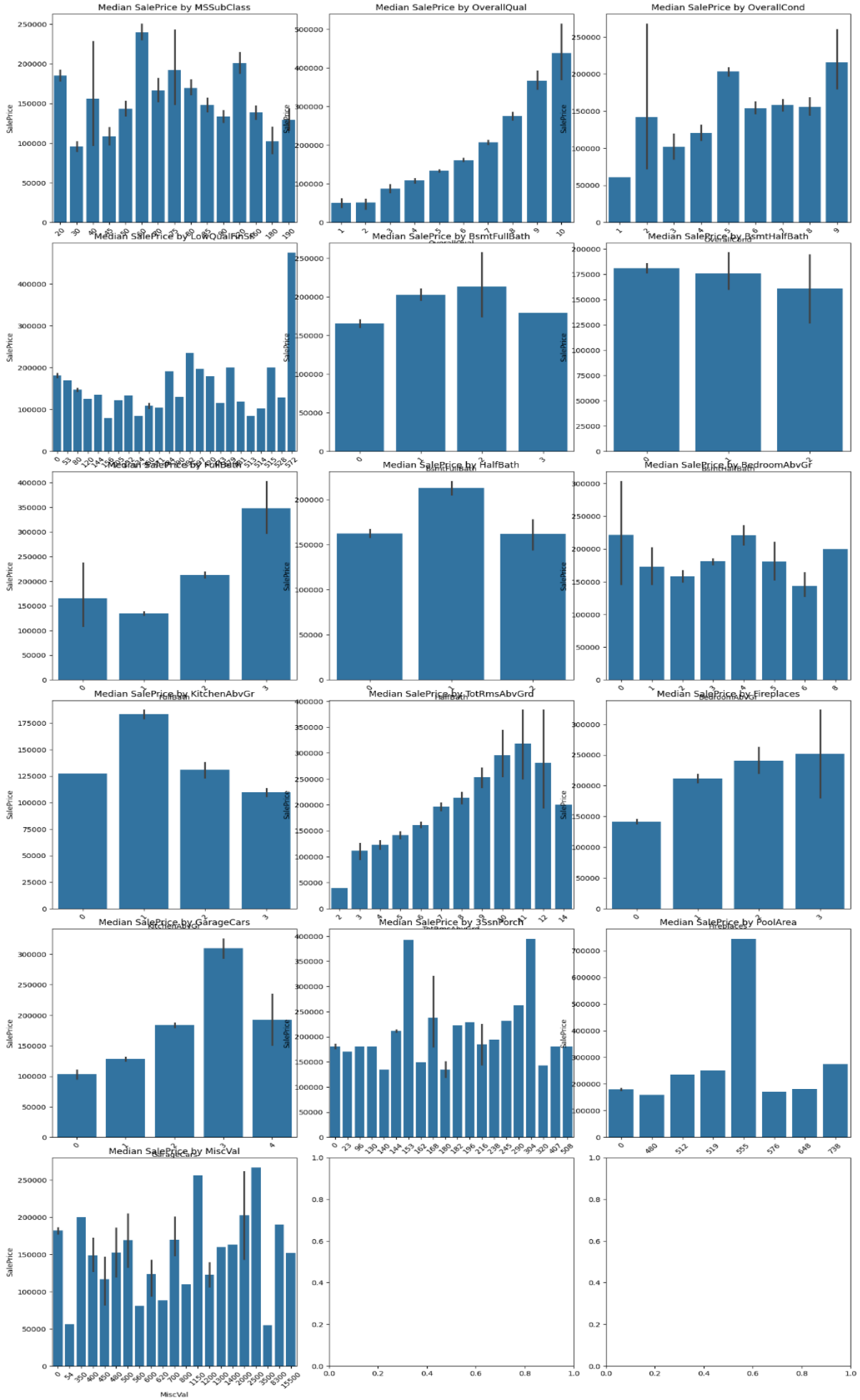

Histogram Distribution of Sale Price
Count
SalePrice
Boxplot Distribution of Sale Price
SalePrice

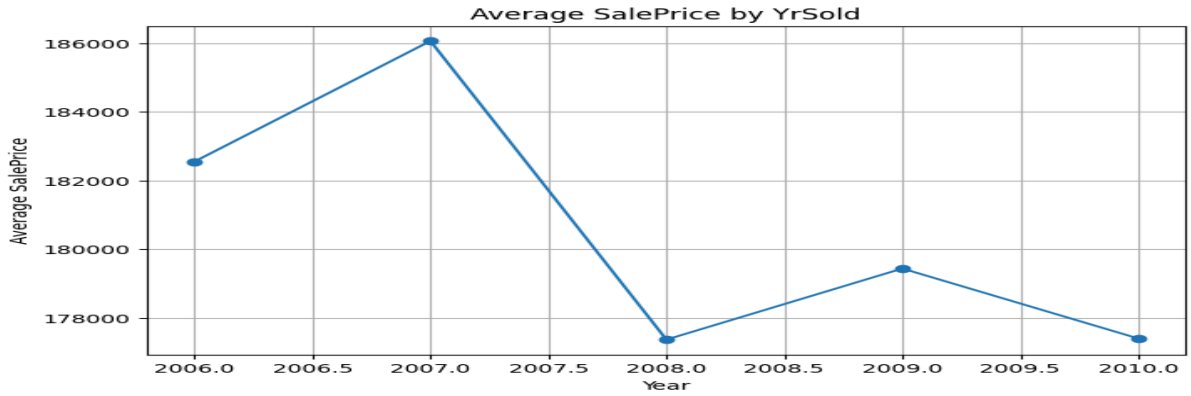

LotFrontage
LotArea
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
GrLivArea
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
ScreenPorch
SalePrice

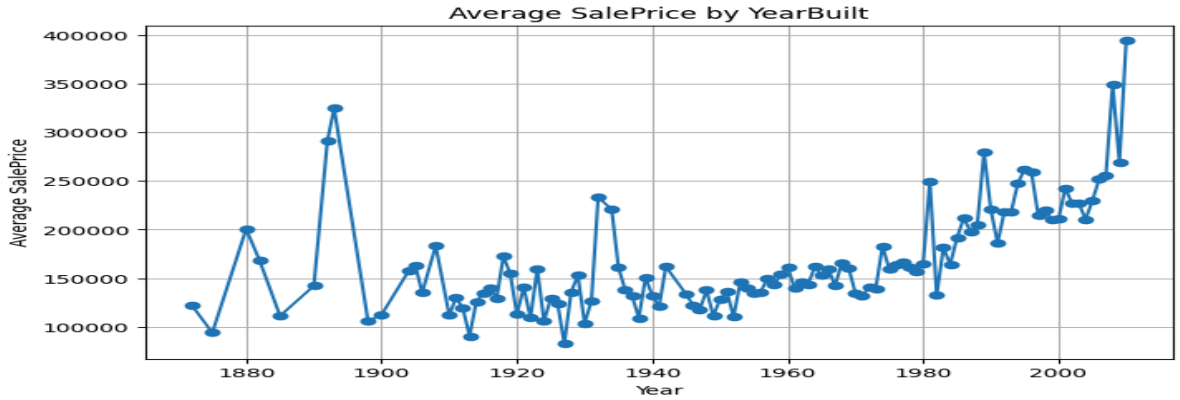
2. *Bivariate analysis*

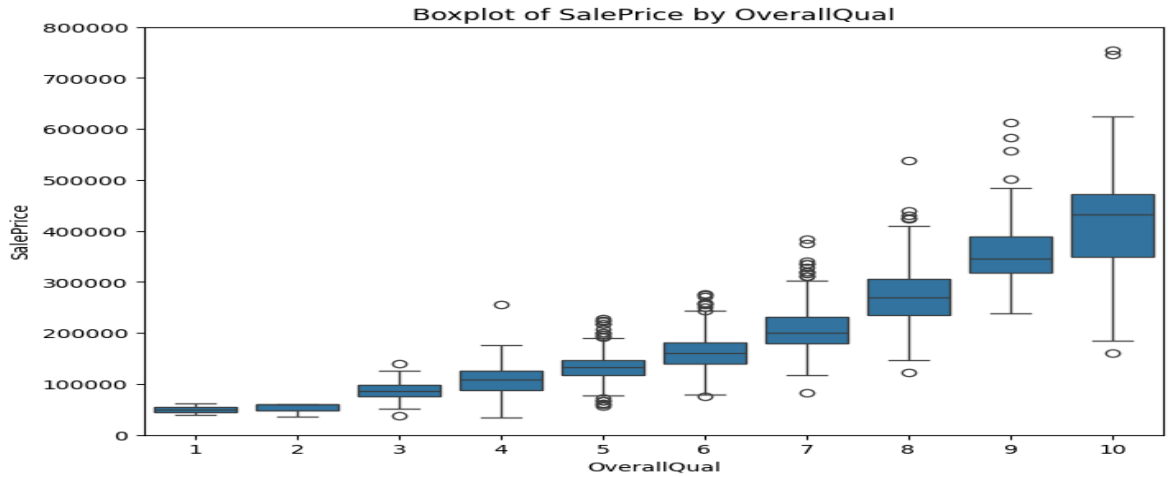
Regression plots, bar plots and correlation matrices were utilised to explore relationships between numerical variables and their correlation with the target variable (house price).




Median SalePrice by MSSubClass
Median SalePrice by OverallQual
Median SalePrice by OverallCond
Median SalePrice by LowQualFinSF
Median SalePrice by BsmtFullBath
Median SalePrice by BsmtHalfBath
Median SalePrice by FullBath
Median SalePrice by HalfBath
Median SalePrice by BedroomAbvGr
Median SalePrice by KitchenAbvGr
Median SalePrice by TotRmsAbvGrd
Median SalePrice by Fireplaces
Median SalePrice by GarageCars
Median SalePrice by 3SsnPorch
Median SalePrice by PoolArea
Median SalePrice by MiscVal
SalePrice
MiscVal

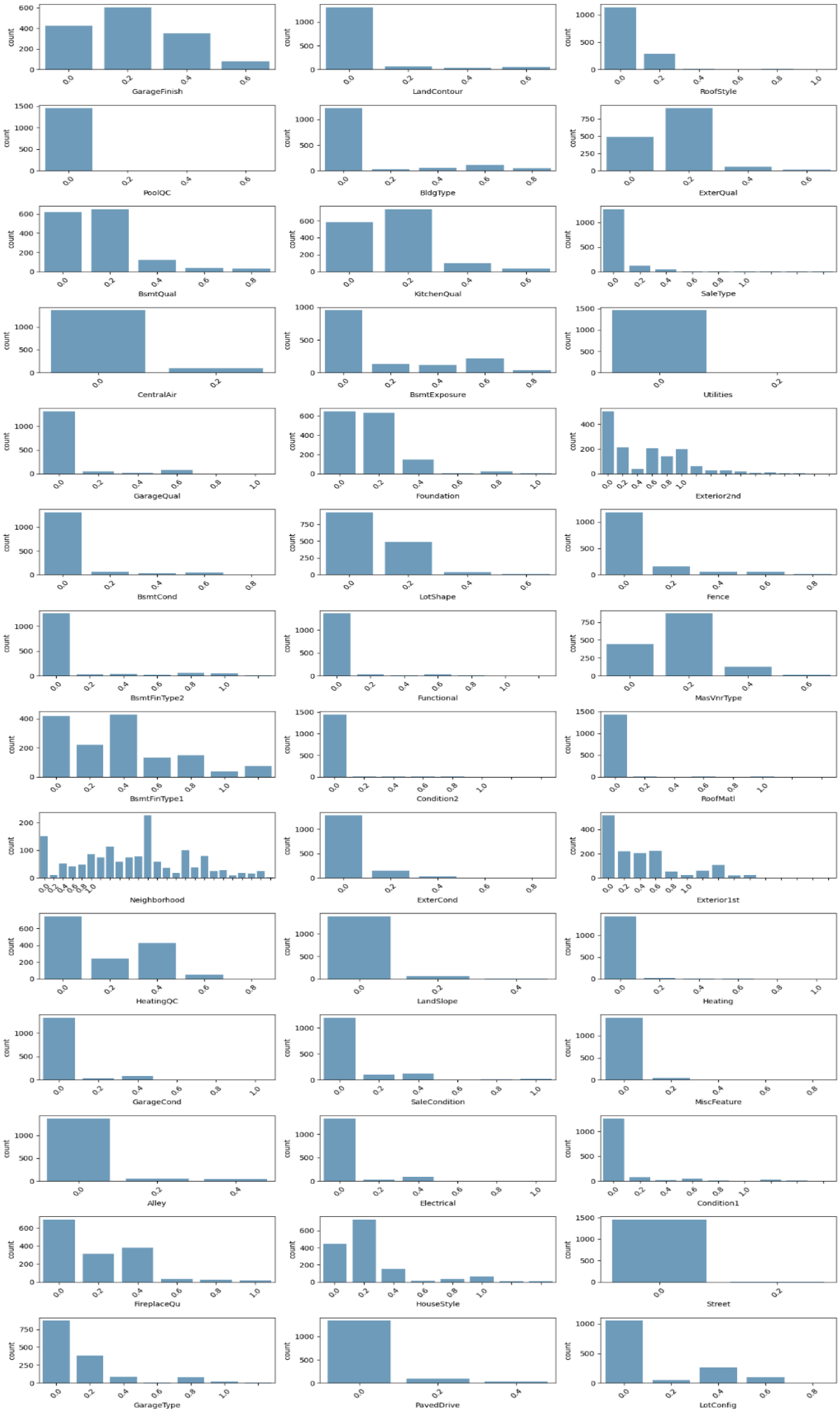






3. *Categorical analysis*

Bar charts and count plots were employed to visualise the distribution of categorical features and their impact on house prices.


count
GarageFinish
LandContour
RoofStyle
PoolQC
BldgType
ExterQual
BsmtQual
KitchenQual
SaleType
CentralAir
BsmtExposure
Utilities
GarageQual
Foundation
Exterior2nd
BsmtCond
LotShape
Fence
BsmtFinType2
Functional
MasVnrType
BsmtFinType1
Condition2
RoofMatl
Neighborhood
ExterCond
Exterior1st
HeatingQC
LandSlope
Heating
GarageCond
SaleCondition
MiscFeature
Alley
Electrical
Condition1
FireplaceQu
HouseStyle
Street
GarageType
PavedDrive
LotConfig

These visualisations provided valuable information about the data's characteristics and guided the feature engineering process.

C. Algorithm and Techniques

Several machine learning algorithms were considered for this project, each with its strengths and weaknesses.

Linear Regression: A baseline model to establish a benchmark for comparison.

Decision Trees: Capable of capturing complex nonlinear relationships between features and the target variable.

Random Forest: An ensemble method that combines multiple decision trees to improve predictive accuracy and reduce overfitting.

Gradient Boosting: An iterative approach that builds a strong predictive model by combining weak learners.

The choice of algorithm was influenced by factors such as model interpretability, computational efficiency, and predictive performance.

D. Benchmark

To establish a baseline for model performance, a simple linear regression model was initially implemented. This model serves as a benchmark against which the performance of more complex models can be compared.

The linear regression model provides a foundation for understanding the relationship between house prices and key features. However, it is expected that more sophisticated models will outperform the baseline due to the complex nature of house price determination.

III. METHODOLOGY

A. Data Preprocessing

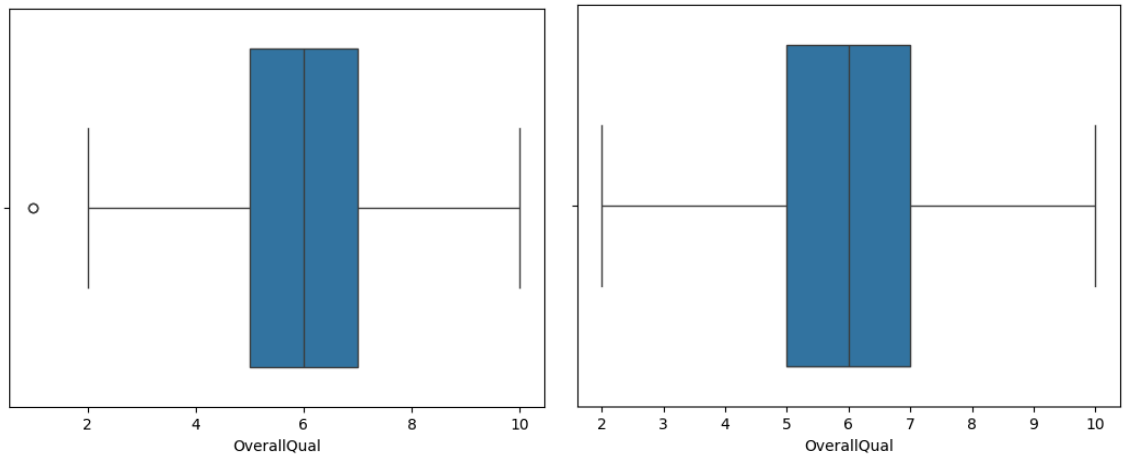
1. Data Cleaning

The first step of data preprocessing was to drop the unnecessary ‘Id’ feature then to check for missing values and handle them. A dataframe containing the sum and percentages of null values in each column was created as well as a heatmap plotted. It was observed that about 19 columns had missing values with certain columns missing almost entire values. Upon further inspection these columns were noted to contain “special features” like a pool, alley or fence which when absent would not denote a house of low quality.

Feature	**Sum**	**Percentage**(%)
PoolQC	1453	99.5

MiscFeature	1406	96.3
Alley	1369	93.8
Fence	1179	80.77

Median imputation was carried out on the continuous columns with missing values due to the presence of outliers. For the categorical columns with missing values, ‘None’ was used to fill null values due to the recognition noted earlier. Outliers were identified using boxplots and addressed through capping and flooring using the interquartile range.



2. Feature engineering

New features were created by combining existing features or extracting relevant information. For example, the age of the house which was derived when the year the house was sold was subtracted from the year the house was built as well as a new column which checked for the presence of a garage. These additional columns would provide additional understanding to the model.

Continuous numerical features were scaled using a Standard Scaler to ensure comparable ranges and prevent bias in model training. Categorical variables were converted into numerical representations using label encoding for ordinal columns and one-hot encoding for the others.

These preprocessing steps altogether aimed to enhance data quality and improve model performance.

B. Implementation

Pre-Modelling

The preprocessed data was split into the training set and the validation set in an 80:20 ratio.[8] The validation set would then be used to test for overfitting i.e. when the model memorises a pattern in the training data but cannot make accurate predictions in what it has not yet reviewed [9] and underfitting i.e. when the model is too simple.

Model Selection and Training

Based on the exploratory data analysis and the nature of the problem, three of the machine learning models initially selected were implemented. Linear Regression as a baseline model for comparison, the Decision Tree Regressor to capture non-linear relationships between features and the target variable and the Random Forest Regressor to improve predictive accuracy through ensemble learning.

The models were trained on 80% of the preprocessed training dataset using appropriate hyperparameters. Grid search or randomised search cross-validation was employed to optimise model performance.

C. Refinement

Hyperparameter Tuning for Random Forest

Model performance was evaluated based on the chosen metrics. Grid search cross-validation was employed to optimise the hyperparameters of the Random Forest model. The following hyperparameter grid was explored:

```
param_grid = {
   'n_estimators': [100, 200, 300],
   'max_depth': [5, 10, 15]
}
```

The grid search identified the following combination of hyperparameters as the most optimal:

- max_depth = 15
- N_estimators = 300

This configuration resulted in the best performance metrics for the Random Forest model.

IV. RESULTS

A. Model Evaluation and Validation

The performance of the trained models was evaluated using the previously defined metrics (MSE, RMSE and R-squared).

Model	MSE	RMSE	MAE	R-squared
Linear Regression	294046.37	542.26	30494.83	-79818915184.09
Decision Tree Regressor	0.48	0.69	0.34	0.78
Random Forest Regressor	0.31	0.56	0.23	0.91

1. Linear Regression

The extremely high RMSE and negative R-squared value indicate a poor model fit. This is likely due to the complex nature of the house price prediction problem and the limitations of linear models in capturing non-linear relationships.

2. Decision Tree Regressor

While showing improvement over linear regression, the decision tree model still exhibits some error. This could be attributed to overfitting or the inherent limitations of decision trees in handling complex patterns.

3. *Random Forest Regressor*

The random forest model demonstrates the best performance among the evaluated models, with a significantly lower RMSE and higher R-squared value. This indicates its ability to capture complex relationships and generalise well to unseen data.

B. Justification

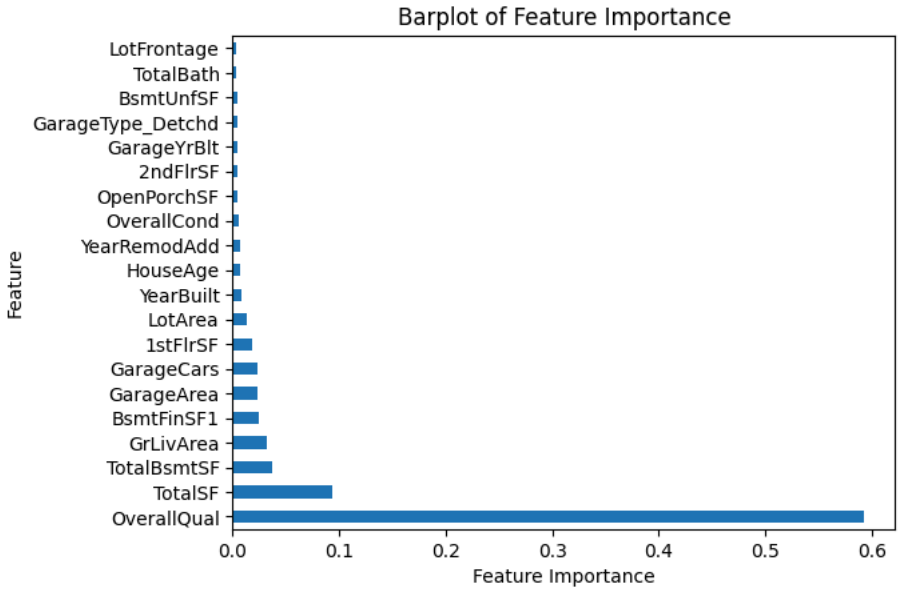
Based on the evaluation metrics, the **Random Forest Regressor** is selected as the preferred model for predicting house prices. It consistently outperforms the other models in terms of accuracy and generalizability. The ensemble nature of random forests allows it to handle complex patterns and reduce overfitting effectively.

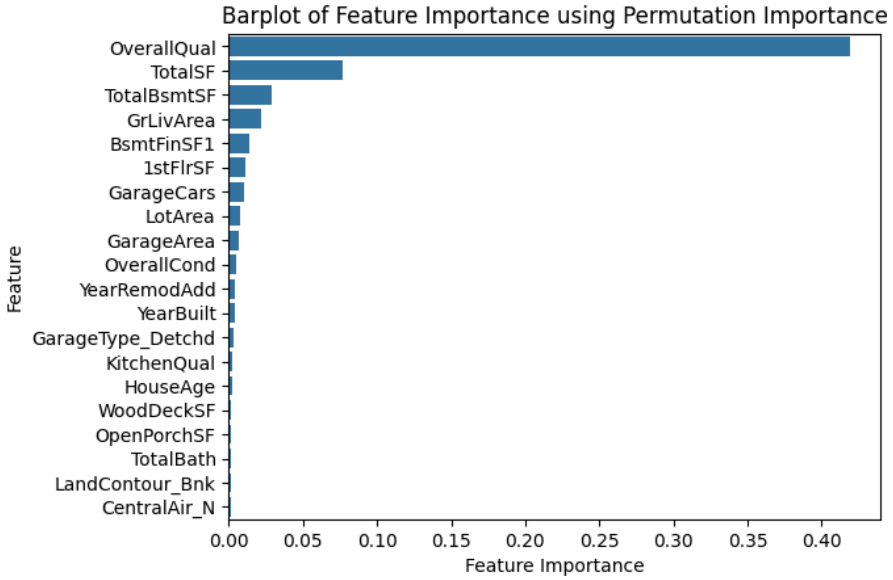
The best mean cross-validation score of 0.882 and test-set score of 0.915 for the Random Forest model indicate good predictive performance and generalisation ability.

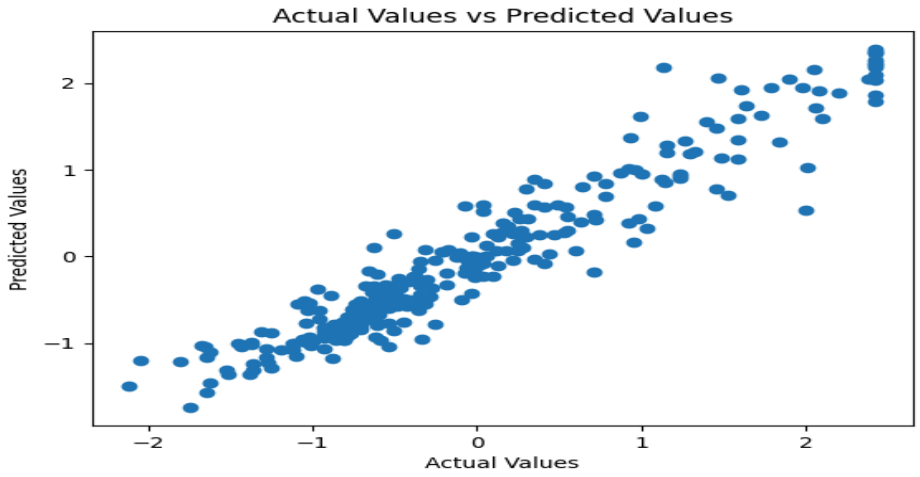
The optimal hyperparameters for the Random Forest model are max_depth = 15 and n_estimators = 300, suggesting that a deeper tree structure with a larger number of estimators improves performance.

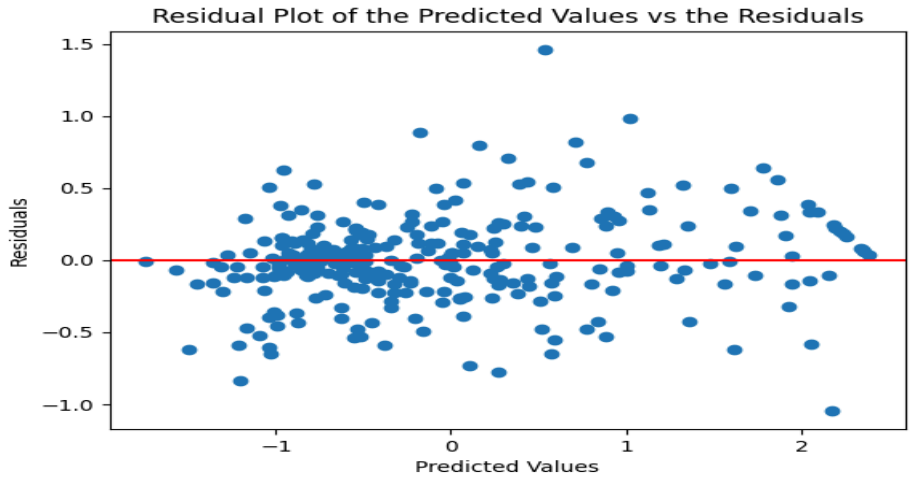
V. CONCLUSION

A. Free-form Visualisation









B. Reflection

The developed Random Forest model demonstrated promising performance in predicting house prices. The model effectively captured complex relationships between features and the target variable, leading to accurate predictions. However, further improvements can be achieved through refined feature engineering, exploring different algorithms, and incorporating larger datasets.

C. Improvement

Potential areas for future model enhancement include incorporating advanced techniques like gradient boosting or neural networks, expanding feature engineering through domain expertise and advanced feature selection, addressing imbalanced data, capturing temporal trends with time series analysis, uncovering hidden patterns through interaction effect exploration, and refining feature engineering for optimal predictive power.

VI. REFERENCES

1. Investopedia. (2023, August 12). What You Should Know About Real Estate Valuation. https://www.investopedia.com/articles/mortgages-real-estate/11/valuing-real-estate.asp
2. ebruiserisobay. (n.d.). House Price Prediction - EDA, ML, End-to-End. Kaggle
3. Vipin20. (n.d.). House Prices EDA & Feature Engineering. Kaggle
4. pawestobrawe. (n.d.). House Prices Advanced Regression Techniques. Kaggle
5. Will Koehrsen. (n.d.). Intro to Model Tuning: Grid and Random Search. Kaggle
6. Sourav Chatterjee. (2021, May 21). Know the Best Evaluation Metrics for Your Regression Model. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/
7. Kaggle. (n.d.). House Prices: Advanced Regression Techniques. Kaggle
8. H2O.ai. (n.d.). Validation Sets. h2o.ai
9. LinkedIn. (n.d.). What Is the Purpose of a Validation Set in Machine Learning? Linkedin
10. krishnaik06. (n.d.). Advanced House Price Prediction. [GitHub repository]. https://github.com/krishnaik06/Advanced-House-Price-Prediction-/blob/master/Feature%20Engineering.ipynb
11. Flavio Henrique CBC. (n.d.). Machine Learning Capstone Project: Final Report. GitHub