House Price Prediction - Project Report

1. Problem Statement

The goal of this project is to predict the sale price of houses based on various features such as size, location, number of bathrooms, year built, and more. Accurate prediction of house prices helps buyers, sellers, and real estate agents make informed decisions.

2. Dataset Overview

The dataset used in this project is from Kaggle's 'House Prices: Advanced Regression Techniques' competition. It contains 1460 observations and 81 features including numeric and categorical variables.

3. Data Cleaning

Missing values were identified and handled appropriately. Columns with too many missing values like 'Alley', 'PoolQC', 'Fence', etc. were dropped. Numeric columns were filled with median values and categorical columns with mode.

4. Exploratory Data Analysis (EDA)

EDA was conducted using visualizations like heatmaps and histograms to understand the correlation between features and SalePrice. 'OverallQual', 'GrLivArea', and 'GarageCars' showed strong positive correlations with SalePrice.

5. Feature Engineering

A subset of top correlated features was selected: OverallQual, GrLivArea, GarageCars, TotalBsmtSF, FullBath, and YearBuilt. Features were scaled using StandardScaler before model training.

6. Model Building & Evaluation

Linear Regression, Ridge Regression, and Lasso Regression models were built. The models were evaluated using MAE, MSE, RMSE, and R2 score. Linear Regression gave good baseline performance with R2 around 0.79.

House Price Prediction - Project Report

7. Conclusion & Insights

The model can predict house prices reasonably well. OverallQual and GrLivArea were found to be the most important features. A Streamlit app was also created to allow users to predict prices interactively.

8. Tools & Libraries Used

Python

Pandas, NumPy

Matplotlib, Seaborn

Scikit-learn

Streamlit

Jupyter Notebook