House Price Prediction Report

1. Introduction

This report presents the house price prediction project using the Ames Housing dataset. The goal was to build a regression model that accurately predicts house prices based on various property features. The project leveraged feature engineering, regression techniques, and model tuning to improve accuracy.

2. Dataset Overview

- **Dataset:** Ames Housing dataset
- **Source:** Public real estate dataset containing **over 2,500 house sales**
- **Target Variable:** Sale Price
- **Number of Features:** 80 (categorical & numerical)

2.1 Key Features Used

- **Lot Area** (Size of the property)
- **Overall Quality** (Construction & material quality)
- **Total Basement Area** (Total size of the basement)
- **Garage Cars** (Number of garage spaces)
- **Year Built** (Year the house was built)
- **Neighborhood** (Location of the house)

3. Data Cleaning & Preprocessing

3.1 Handling Missing Data

- **Imputed missing values** using mean/median for numerical features
- **Filled categorical missing values** with 'Umknown' or mode

3.2 Feature Engineering

- Created new features (e.g., Total Square Footage = Basement + Ground Floor Area)
- Converted categorical variables using One-Hot Encoding
- Removed highly correlated features to reduce multicollinearity

4. Model Selection & Training

4.1 Regression Models Tested

Model	Initial R^2 Score	Optimized R^2 Score
Linear Regression	0.85	0.89
Ridge Regression	0.84	0.88

4.2 Optimization Techniques

- Regularization (Ridge Regression to improve generalization)
- Feature Selection (Kept 50 best features instead of 80)
- Scaling (Standardized numerical features to improve performance)

5. Results & Insights

5.1 Final Model Performance

- **Best Model:** Ridge Regression with R^2 = **0.89**
- **Feature Importance Analysis:** Removing low-impact features increased model efficiency.

5.2 Observations

- **Newer homes** tend to have higher prices.

- **Location (Neighborhood)** significantly impacts property value.
- **Garage space** has a moderate effect on pricing.

6. Challenges & Solutions

Challenges Faced:

- **Overfitting:** Too many features led to poor generalization.
- **Multicollinearity:** Correlated features affected model stability.
- **Skewed Data:** Some features had high skewness, affecting predictions.

Solutions Implemented:

- ❖ **Feature Reduction** (Kept only relevant variables)
- ❖ **Regularization (Ridge/Lasso)** to prevent overfitting
- ❖ **Log Transformation** for skewed features like Sale Price

7. Conclusion

This project successfully predicted house prices using **machine learning techniques**, improving the model from **R^2 = 0.85 to 0.89** through feature engineering and optimization.