BITS PILANI DIGITAL

FIRST TRIMESTER 2025-26

ADVANCED APEX PROJECT 1

Project Title	**Real Estate Price Prediction**
Supervisor Name	Bharathi Dasari
Name of the Learner (with BITS ID)	

Name of the Learner	BITS ID
Anik Das	2025EM1100026
Adeetya Wadikar	2025EM1100384
Tushar Nishane	2025EM1100306

Courses Relevant for the Project & Corresponding Trimester

Sl. No.	Subject Name	State the relevance to Project
1	Statistical Modelling & Inferencing	• Simple Linear Regression (Overall Quality predictor, R^2 = 0.6325 train, R^2 = 0.6512 test, RMSE = \$52,879) • Multiple Linear Regression (73 features, R^2 = 0.8609 train, R^2 = 0.8489 test, RMSE = \$34,807) • Ridge Regression with L2 regularization (α = 1.0, R^2 = 0.8606 train, R^2 = 0.8494 test, RMSE = \$34,746) • K-Fold Cross-Validation (5-fold, Mean R^2 = 0.8479 ± 0.0242) • VIF analysis, Residual analysis, Q-Q plots, Shapiro-Wilk test
2	Data Pre-processing	• Missing value analysis: 15,749 values in 27 columns • 4-step imputation strategy: - Drop columns >50% missing (5 columns) - Fill 'None' for categorical (10 columns) - Median imputation (Lot Frontage → 68, Garage Yr Blt → 1979) - Mode imputation (Electrical → 'SBrkr') • Duplicate detection: 0 duplicates found • Low variance removal: 6 columns dropped • Train-Test split: 80/20 (2,344/586 samples)
3	Feature Engineering	• Created 5 domain-based features: - Total_SF = Gr Liv Area + Total Bsmt SF (r = 0.79) - Total_Bathrooms, Total_Porch_SF, House_Age, Years_Since_Remod • Log1p transformation for 21 skewed features (skewness > 1) • Label Encoding for 32 categorical features • Feature Selection using Random Forest Importance (Top: Overall Qual = 0.48, Total_SF = 0.31) • Multicollinearity: Garage Cars ↔ Garage Area (r = 0.89)
4	Data Visualization & Storytelling	• SalePrice distribution histogram (skewness = 1.74) • Box plots, Missing value bar chart (missingno library) • Correlation heatmap (top 12 features), Pair plots • Scatter plots: 6 features vs SalePrice • Model comparison bar charts, Actual vs Predicted plot • Q-Q plot, Residual distribution histogram • Sale Price by Quality Rating, Price vs Area by Neighborhood • Feature Importance (Top 20), Project Summary Dashboard
5	Data Stores & Pipelines	• CSV data loading using Pandas read_csv • Directory structure: data/, docs/, notebooks/ • Data versioning: Raw (AmesHousing.csv), Cleaned (AmesHousing_cleaned.csv), Engineered (AmesHousing_engineered.csv) • Schema documentation: schema_summary.csv • Cross-platform path handling using pathlib • Memory: 6.92 MB for 2,930 × 82 dataset • Reproducible Jupyter Notebook workflow

1. PROBLEM STATEMENT

Accurate real estate valuation is essential for buyers, sellers, and financial institutions. Traditional valuation methods can be subjective and time-consuming. This project develops machine learning models to predict house sale prices objectively based on property characteristics.

Key Challenges:

- Overpricing or underpricing of properties
- Inefficient negotiation processes
- Poor investment decisions
- Lack of transparency in property valuation

2. BUSINESS GOAL

Primary Objective:

Develop a predictive regression model that estimates residential property sale prices with high accuracy. The model should help stakeholders:

- **Buyers:** Assess fair market value before purchase
- **Sellers:** Set competitive listing prices
- **Investors:** Identify undervalued properties
- **Lenders:** Support loan underwriting decisions

Success Criteria vs Achieved Results:

Metric	Target	Achieved	Status
R-squared (R^2)	> 0.80	0.8494 (84.94%)	✓
RMSE	< 15% of avg price	$34,746	✓
MAE	-	$21,557	✓
Overfitting Gap (Train-Test)	< 0.03	0.0112	✓

MODEL PERFORMANCE SUMMARY

Dataset Size:	2,930 properties × 82 features	**R-squared (Test):**	0.8494 (84.94%)
Final Features Used:	73	**RMSE:**	$34,746
Best Model:	Ridge Regression (α = 1.0)	**Top Predictor:**	Overall Quality (r = 0.80)

3. DATA SOURCE

Dataset Name:	Ames Housing Dataset
Platform:	Kaggle
URL:	https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset
Original Source:	Ames, Iowa Assessor's Office (compiled by Dean De Cock)

Dataset Specifications:

Specification	Value
File Format	CSV (AmesHousing.csv)
Total Records	2,930 residential properties
Total Features	82 columns (including target variable)
Target Variable	SalePrice (continuous, USD)
Time Period	Properties sold in Ames, Iowa (2006-2010)
Memory Usage	6.92 MB

Feature Breakdown:

- Numerical (int64): 23 columns
- Numerical (float64): 16 columns
- Categorical (object): 43 columns

Target Variable (SalePrice) Statistics:

Minimum:	$12,789
Maximum:	$755,000
Mean:	$180,796
Median:	$160,000
Std Dev:	$79,887

4. TOOLS & TECHNOLOGIES

Category	Tools / Technologies
Programming Language	Python 3.12
Development Environment	Jupyter Notebook, VS Code
Data Manipulation	Pandas 2.2.3, NumPy 1.26.4
Machine Learning	Scikit-learn (LinearRegression, Ridge, RidgeCV,
	RandomForestRegressor, StandardScaler, LabelEncoder,
	KFold, train_test_split, metrics)
Statistical Analysis	SciPy (stats: shapiro, probplot, skew)
	Statsmodels (variance_inflation_factor)
Data Visualization	Matplotlib, Seaborn, Missingno
Version Control	Git, GitHub
Operating System	Linux (Ubuntu)

5. PROJECT WORKFLOW

Phase 1: Data Acquisition

■■■ 1.1 Environment Setup (import libraries)
■■■ 1.2 Data Loading (CSV from Kaggle)
■■■ 1.3 Initial Data Inspection (shape, dtypes, info)
■■■ 1.4 Schema Validation (column verification)
■■■ 1.5 Data Quality Assessment (missing values, duplicates, target stats)

Phase 2: Preprocessing & Exploratory Analysis

■■■ 2.1 Summary Statistics (descriptive stats, target analysis)
■■■ 2.2 Missing Value Analysis (27 columns with missing data)
■■■ 2.3 Missing Value Treatment (4-step strategy)
■■■ 2.4 Univariate Analysis (distributions, skewness)
■■■ 2.5 Low-Variance Feature Removal (6 columns dropped)
■■■ 2.6 Bivariate Analysis & Correlations (heatmap, scatter plots)
■■■ 2.7 Outlier Detection (IQR method)

Phase 3: Feature Engineering

■■■ 3.1 Feature Creation (5 new features)
■■■ 3.2 Skewness Handling (log1p transformation)
■■■ 3.3 Categorical Encoding (Label Encoding)
■■■ 3.4 Feature Importance Analysis (Random Forest)

Phase 4: Modeling & Evaluation

■■■ 4.1 Data Preparation (80/20 train-test split)
■■■ 4.2 Simple Linear Regression (baseline)
■■■ 4.3 Multiple Linear Regression
■■■ 4.4 Ridge Regression (L2 regularization with cross-validation)
■■■ 4.5 Model Comparison & Selection
■■■ 4.5.1 Residual Diagnostics (Q-Q plot, Shapiro-Wilk test)

Phase 5: Visualization & Storytelling

■■■ 5.1 Dashboard Visualizations
■■■ 5.2 Key Insights
■■■ 5.3 Conclusions & Recommendations

6. DATA EXTRACTION

Source Platform:	Kaggle
Dataset URL:	https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset
Download Method:	Manual download from Kaggle
File Downloaded:	AmesHousing.csv
Storage Location:	/data/AmesHousing.csv

Data Loading Code Used:

```
from pathlib import Path
import pandas as pd

notebook_dir = Path().resolve()
data_path = notebook_dir.parent / 'data' / 'AmesHousing.csv'
df = pd.read_csv(data_path)
```

Verification Output:

- Dataset Dimensions: 2,930 rows × 82 columns
- Memory Usage: 6.92 MB
- Duplicate Rows: 0
- Target Variable (SalePrice) Missing: 0

7. SCHEMA / DATA DICTIONARY

DATASET OVERVIEW:

Total Columns: 82 • Numerical (int64): 23 • Numerical (float64): 16 • Categorical (object): 43

TARGET VARIABLE:

Column: SalePrice | Type: int64 | Description: Property sale price in USD
Statistics: Min: $12,789 | Max: $755,000 | Mean: $180,796 | Median: $160,000 | Std Dev: $79,887

KEY NUMERICAL FEATURES (Top Correlations with SalePrice):

Feature	Correlation	Description
Overall Qual	0.80	Overall material and finish quality (1-10)
Gr Liv Area	0.71	Above grade living area (sq ft)
Garage Cars	0.65	Garage capacity (number of cars)
Garage Area	0.64	Garage size (sq ft)
Total Bsmt SF	0.63	Total basement area (sq ft)
1st Flr SF	0.62	First floor area (sq ft)
Year Built	0.56	Original construction year
Full Bath	0.55	Full bathrooms above grade
Year Remod/Add	0.53	Remodel year
Garage Yr Blt	0.51	Year garage was built
Mas Vnr Area	0.50	Masonry veneer area (sq ft)
TotRms AbvGrd	0.50	Total rooms above grade
Fireplaces	0.47	Number of fireplaces

KEY CATEGORICAL FEATURES:

Feature	Unique Values	Description
Neighborhood	28	Physical location within Ames
MS Zoning	7	Zoning classification
Bldg Type	5	Type of dwelling
House Style	8	Style of dwelling
Exterior 1st	16	Exterior covering on house
Foundation	6	Type of foundation
Heating QC	5	Heating quality and condition
Central Air	2	Central air conditioning (Y/N)
Garage Type	7	Garage location
Sale Condition	6	Condition of sale

7. SCHEMA / DATA DICTIONARY (Continued)

MISSING VALUE ANALYSIS:

Total Missing Values: 15,749 | Columns with Missing: 27 out of 82

Feature	Missing Count	Missing %
Pool QC	2,917	99.56%
Misc Feature	2,824	96.38%
Alley	2,732	93.24%
Fence	2,358	80.48%
Mas Vnr Type	1,775	60.58%
Fireplace Qu	1,422	48.53%
Lot Frontage	490	16.72%

COLUMNS DROPPED (>50% Missing):

Pool QC (99.56%) • Misc Feature (96.38%) • Alley (93.24%) • Fence (80.48%) • Mas Vnr Type (60.58%)

LOW VARIANCE COLUMNS DROPPED (6 columns):

Street (99.6% Pave) • Utilities (99.9% AllPub) • Condition 2 (99.0% Norm)
Roof Matl (98.5% CompShg) • Heating (98.5% GasA) • Land Slope (95.2% Gtl)

ENGINEERED FEATURES (Created in Phase 3):

Feature	Formula	Correlation
Total_SF	Gr Liv Area + Total Bsmt SF	0.79
Total_Bathrooms	Full Bath + 0.5*Half Bath + Bsmt Baths	0.64
Total_Porch_SF	Sum of all porch areas	0.38
House_Age	Yr Sold - Year Built	-0.56
Years_Since_Remod	Yr Sold - Year Remod/Add	-0.53

FINAL DATASET AFTER PREPROCESSING:

Original: 2,930 × 82 → After high-missing drop: 2,930 × 77 → After low-variance drop: 2,930 × 71
After engineered features: 2,930 × 76 | Features for modeling: 73 | Train: 2,344 (80%) | Test: 586 (20%)

8. MODEL PERFORMANCE SUMMARY

Model	Features	R^2 (Train)	R^2 (Test)	RMSE	MAE	Gap
Simple Linear Reg	1	0.6325	0.6512	$52,879	$36,141	0.0187
Multiple Linear Reg	73	0.8609	0.8489	$34,807	$21,622	0.0120
Ridge Reg (α=1.0)	73	0.8606	0.8494	$34,746	$21,557	0.0112

Cross-Validation Results (5-Fold, Ridge):

R^2 scores per fold: [0.8494, 0.8161, 0.8607, 0.8284, 0.8849]
Mean R^2: 0.8479 ± 0.0242 | Mean RMSE: $30,908 ± $3,054

Selected Model: Ridge Regression

Reason: Lowest overfitting gap (0.0112), handles multicollinearity via L2 regularization