

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	Descriptive statistics, correlation analysis, distribution analysis, feature importance ranking (Random Forest), statistical validation of engineered features. Applied in EDA phase to identify relationships between features and target variable (SalePrice).
2	Data Pre-processing	Missing value imputation (median, mode, zero-fill), outlier detection using IQR method, duplicate removal, low-variance feature removal, data type validation. Cleaned dataset from 27 columns with missing values to 99.9% completeness.
3	Feature Engineering	Created 13 new features (aggregate, temporal, binary, interaction), applied log transformations for skewed features, categorical encoding (ordinal & label encoding), correlation-based feature selection, dimensionality reduction from 82 to 69 features.
4	Data Visualization & Storytelling	Univariate analysis (39 histograms, 33 count plots), bivariate analysis (scatterplots, boxplots), multivariate analysis (correlation heatmaps). Created comprehensive visualizations to identify patterns and relationships in the housing data.
5	Data Stores & Pipelines	CSV data extraction, reproducible data loading from Kaggle dataset, data schema validation, cross-validation with data dictionary, engineered dataset persistence (AmesHousing_engineered.csv). Maintained data lineage throughout preprocessing pipeline.

Project Title: Real Estate Price Prediction

1. Problem Statement

Accurate real estate price prediction is a major challenge for buyers, sellers, and investors. Traditional valuation methods rely on limited appraisal factors and often fail to capture the diverse influences such as lot size, number of rooms, construction year, location, and neighborhood characteristics. Inaccurate estimates may lead to poor investment decisions, buyer dissatisfaction, or overpricing in competitive markets. We need a data-driven approach to analyze historical housing data and build predictive models that can estimate house sale prices with greater reliability. This solution will also help identify which features most strongly affect property prices.

2. Business Goal

The primary goal is to develop a machine learning regression model that predicts property sale prices with strong accuracy and reliability. This will allow potential buyers and investors to make informed decisions, real estate agencies to improve their valuation practices, and stakeholders to understand the market drivers of property prices. By the end of the project, we aim to deliver a working predictive model along with clear insights and visualizations.

3. Data Source

We will use the **Ames Housing Dataset**, a widely used benchmark dataset for real estate price prediction.

Source Platform: Kaggle

Full Citation: Shashank Necrothapa. (n.d.). Ames Housing Dataset [Data set]. Kaggle. Retrieved October 1, 2025, from <https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset>

Dataset Details:

- **Instances:** 2,930 observations
- **Features:** 82 attributes (numerical and categorical)
- **Target:** SalePrice (house sale price)
- **Features describe:** lot size, rooms, year built, neighborhoods, etc.

4. Tools & Technologies

- **Programming Language:** Python 3.12
- **Core Libraries:** Pandas (2.2.3), NumPy (1.26.4)
- **Machine Learning:** Scikit-learn (RandomForestRegressor, LabelEncoder, StandardScaler, feature selection tools)
- **Data Visualization:** Matplotlib, Seaborn, Missingno
- **Development Environment:** Jupyter Notebook
- **BI Tools (Phase 4):** Tableau or Power BI for final dashboard
- **Version Control:** GitHub

5. Project Workflow

The project will follow the standard data science lifecycle:

- **Data Acquisition:** Obtain the dataset from Kaggle (direct download method).
- **Preprocessing:** Handle missing values, clean inconsistencies, encode categorical variables.
- **Exploratory Data Analysis (EDA):** Use statistical summaries and visualizations to study patterns.
- **Feature Engineering:** Create new features (e.g., house age, price per sqft) and perform scaling/encoding.
- **Model Building:** Train regression models (Linear Regression, Ridge/Lasso, Random Forest, Gradient Boosting).
- **Model Evaluation:** Evaluate using RMSE, MAE, and R².
- **Visualization & Reporting:** Create dashboards and storytelling slides highlighting findings.

6. Data Extraction

The dataset (Ames Housing) has been acquired using the direct download method from Kaggle. We manually downloaded the ZIP file from the Kaggle dataset page and extracted it into the project folder for further preprocessing.

Files: AmesHousing.csv

Location: ./data/

This ensures that our dataset remains consistent with the official Kaggle version. A short inspection notebook will be prepared to confirm schema details and enable reproducibility in later phases.

7. Schema / Data Dictionary

A schema or data dictionary provides a structured overview of the dataset. For Phase 1, we prepared a concise dictionary with 20 key features that are most relevant for modeling house prices. The complete dictionary covering all 82 features of the Ames Housing dataset will be maintained separately in an Excel file (data_dictionary.xlsx) and updated as the project progresses.

Feature	Type	Description	PK
Order	Integer	Observation number	No
PID	Integer	Unique property identifier	Yes
MS SubClass	Categorical	Type of dwelling (e.g., 20 = 1-Story 1946+)	No
MS Zoning	Categorical	General zoning classification	No
Lot Frontage	Integer	Linear feet of street connected to property	No
Lot Area	Integer	Lot size in square feet	No
Street	Categorical	Type of road access (Grvl, Pave)	No
Lot Shape	Categorical	General shape of property	No
Neighborhood	Categorical	Physical location within Ames city	No
Condition 1	Categorical	Proximity to main road or conditions	No
Bldg Type	Categorical	Type of dwelling (1Fam, Twnhse, etc.)	No
House Style	Categorical	House style (1Story, 2Story, etc.)	No
Overall Qual	Integer	Overall material and finish quality (1-10)	No
Overall Cond	Integer	Overall condition rating (1-10)	No
Year Built	Integer	Original construction year	No
Year Remod/Add	Integer	Year of remodel/addition (if any)	No
Gr Liv Area	Integer	Above-grade living area (sq ft)	No
Full Bath	Integer	Full bathrooms above grade	No
Bedroom AbvGr	Integer	Bedrooms above ground level	No
Garage Cars	Integer	Garage capacity (number of cars)	No
Garage Area	Integer	Garage size in square feet	No
SalePrice	Integer	Property sale price (Target variable)	No