Feature Engineering Assignment Report

House Price Prediction Dataset

Comprehensive Feature Engineering Pipeline

Student Name: Anik Das

Student ID: 2025EM1100026

Institution: Birla Institute of Technology and Science, Pilani

Program: BITS Pilani Digital

Course: Feature Engineering

Date: November 5, 2025

\newpage

Executive Summary

This report documents a comprehensive feature engineering pipeline applied to the Ames Housing dataset, transforming 81 raw features with significant missing data into a clean, normalized dataset ready for machine learning. The project demonstrates strategic decision-making in data preprocessing, feature creation, encoding strategies, and dimensionality reduction.

Key Achievements:

- Successfully handled 7,829 missing values across 19 columns using contextual strategies

- Created 10 meaningful engineered features based on domain knowledge

- Developed 3 composite text-based features capturing property descriptions

- Applied appropriate encoding techniques for 43 categorical variables

- Reduced dimensionality from 220 features to 136 principal components while retaining 95% variance

- Integrated and analyzed student-specific random feature throughout the pipeline

Final Deliverables:

1. Cleaned and transformed dataset with zero missing values

2. Feature-engineered dataset with 220 features (processed_data_engineered.csv)

3. Dimensionally-reduced dataset with 136 PCA components (processed_data_pca.csv)

4. Fully documented Jupyter notebook with code, outputs, and analysis

\newpage

1. Introduction

1.1 Assignment Objective

The primary objective of this assignment is to design and execute a comprehensive feature engineering strategy that transforms raw property data into a machine learning-ready dataset. This involves strategic decision-making in:

- Data quality assessment and cleaning
- Missing value treatment strategies
- Feature transformation and creation
- Categorical data encoding
- Text-based feature representation
- Dimensionality reduction
- Documentation of all decisions with clear justifications

1.2 Dataset Overview

Dataset: Ames Housing Dataset
Source: train.csv
Initial Dimensions: 1,460 observations × 81 features
Target Variable: SalePrice (house sale prices in dollars)
Feature Types: 43 categorical, 38 numeric
Price Range: $34,900 - $755,000

1.3 Student-Specific Requirement

As per assignment requirements, a unique random feature was generated based on my student ID:

- **Student ID:** 2025EM1100026
- **Last 7 Digits (ID_last7):** 1100026
- **Random Seed:** 26 (ID_last7 % 1000)

- **Offset:** 4 (ID_last7 % 7)
- **Feature Name:** `student_random_feature`

This feature was integrated into all exploratory analysis, correlation studies, and dimensionality reduction processes.

\newpage

2. Data Understanding & Initial Assessment

2.1 Feature Type Classification

Upon initial exploration, features were categorized as:

Numeric Features (38):
- **Continuous:** LotArea, GrLivArea, TotalBsmtSF, GarageArea, etc.
- **Discrete:** YearBuilt, YearRemodAdd, BedroomAbvGr, FullBath, etc.
- **Student Feature:** student_random_feature (uniform distribution, range: 5-103)

Categorical Features (43):
- **Nominal:** Neighborhood (25 categories), Exterior1st (15 categories), MSZoning (5 categories)
- **Ordinal:** ExterQual, KitchenQual, BsmtQual (quality ratings: Ex > Gd > TA > Fa > Po)
- **Binary:** Street, CentralAir, PavedDrive

2.2 Missing Values Analysis

Total Missing Values: 7,829 cells (6.6% of dataset)
Columns Affected: 19 out of 81 features

Missingness Categories:

Category	Percentage Range	Features	Count
High	>30% missing	PoolQC (99.5%), MiscFeature (96.3%), Alley (93.8%), Fence (80.8%)	4
Moderate	5-30%	LotFrontage (17.7%), FireplaceQu (47.3%), GarageType (5.5%)	7
Low	<5%	BsmtExposure (2.6%), Electrical (0.07%)	8

Key Insight: Most high-missing features represent "absence of feature" rather than missing data (e.g., NA in PoolQC means "no pool").

2.3 Target Variable Analysis

SalePrice Distribution:
- **Mean:** $180,921
- **Median:** $163,000
- **Skewness:** 1.88 (right-skewed)
- **Range:** $34,900 - $755,000

Decision: Apply log transformation to normalize the target variable for better model performance.

\newpage

3. Data Cleaning Strategy

3.1 Missing Value Treatment

Strategic Approach: Context-based imputation rather than blanket strategies.

3.1.1 Categorical Features

Strategy 1: "Absence" Interpretation

For features where NA means "feature doesn't exist":

- **Features:** PoolQC, MiscFeature, Alley, Fence, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2

- **Treatment:** Fill with 'None'

- **Rationale:** Preserves semantic meaning (no pool vs. unknown pool quality)

Strategy 2: Mode Imputation

For true missing values in categorical data:

- **Features:** Electrical, MSZoning, Utilities, Exterior1st, Exterior2nd, SaleType

- **Treatment:** Fill with most frequent category

- **Rationale:** Maintains distribution, appropriate for low missingness

3.1.2 Numeric Features

Strategy 1: Zero Imputation

For features representing areas/quantities where absence means zero:

- **Features:** MasVnrArea, GarageArea, GarageCars, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF

- **Treatment:** Fill with 0

- **Rationale:** Zero accurately represents absence of the feature

Strategy 2: Group-Based Imputation

For contextual numeric features:

- **Feature:** LotFrontage (street frontage)

- **Treatment:** Fill with neighborhood median

- **Rationale:** Lot frontage varies by neighborhood characteristics; group-based imputation preserves spatial patterns

Result: Zero missing values after treatment

3.2 Outlier Detection and Treatment

Method: Interquartile Range (IQR) method

- **Threshold:** Values beyond Q1 - 1.5×IQR or Q3 + 1.5×IQR

- **Identified:** 1 extreme outlier in GrLivArea with very low SalePrice

- **Treatment:** Removed (1,460 → 1,459 observations)

- **Rationale:** Extreme outlier likely represents data entry error or unusual sale circumstance

\newpage

4. Feature Engineering

4.1 Numeric Feature Transformation

Objective: Normalize skewed distributions to improve model performance.

Analysis:

- Calculated skewness for all 38 numeric features
- Threshold: |skewness| > 0.5 indicates significant skew
- **Result:** 29 features identified as highly skewed

Transformation Applied: Log transformation (log1p to handle zeros)

Skewed Features Transformed:

- Area features: LotArea, MasVnrArea, TotalBsmtSF, GrLivArea, GarageArea
- SF features: BsmtFinSF1, 1stFlrSF, 2ndFlrSF, LowQualFinSF
- Porch features: WoodDeckSF, OpenPorchSF, EnclosedPorch, ScreenPorch
- Time features: YearBuilt, YearRemodAdd, GarageYrBlt

Student Random Feature: Not transformed as it follows a uniform distribution (not skewed)

Result: Average skewness reduced from 2.14 to 0.53

4.2 Feature Creation

Strategy: Create domain-relevant features by combining existing attributes.

4.2.1 New Numeric Features Created (10)

Feature Name	Formula	Rationale
TotalSF	TotalBsmtSF + 1stFlrSF + 2ndFlrSF	Total living space is more predictive than individual floors
TotalBath	FullBath + (0.5 × HalfBath)	Aggregated bathroom count
HouseAge	YrSold - YearBuilt	Age at time of sale affects price
RemodAge	YrSold - YearRemodAdd	Time since renovation matters
TotalPorchSF	WoodDeckSF + OpenPorchSF + EnclosedPorch + 3SsnPorch + ScreenPorch	Total outdoor space
HasPool	PoolArea > 0	Binary: pool presence/absence
HasGarage	GarageArea > 0	Binary: garage presence/absence
Has2ndFloor	2ndFlrSF > 0	Binary: two-story indicator
HasBasement	TotalBsmtSF > 0	Binary: basement presence/absence
HasFireplace	Fireplaces > 0	Binary: fireplace presence/absence

Domain Knowledge Applied: Real estate values are influenced by total space, age, and presence of key features (pool, garage, etc.).

4.3 Categorical Feature Encoding

Three-Strategy Approach:

4.3.1 Ordinal Encoding

For features with inherent order (quality/condition ratings):

Mapping Applied:

- **Quality Features:** Ex=5, Gd=4, TA=3, Fa=2, Po=1, None=0

- ExterQual, ExterCond, BsmtQual, BsmtCond, HeatingQC, KitchenQual, FireplaceQu, GarageQual, GarageCond, PoolQC

- **Exposure:** Gd=4, Av=3, Mn=2, No=1, None=0 (BsmtExposure)

- **Basement Finish:** GLQ=6, ALQ=5, BLQ=4, Rec=3, LwQ=2, Unf=1, None=0

- **Functional:** Typ=8, Min1=7, Min2=6, Mod=5, Maj1=4, Maj2=3, Sev=2, Sal=1

Rationale: Preserves ordinal relationships in numeric form.

4.3.2 One-Hot Encoding

For nominal categorical features (no inherent order):

Features Encoded:

- MSZoning (5 categories), Neighborhood (25 categories)

- BldgType (5), HouseStyle (8), RoofStyle (6), RoofMatl (8)

- Exterior1st (15), Exterior2nd (16), Foundation (6)

- Heating (6), Electrical (5), GarageType (6)

- SaleType (9), SaleCondition (6)

- And others...

Result: Created 176 binary indicator columns

4.3.3 Label Encoding

For binary categorical features:

- Street (Pave/Grvl), CentralAir (Y/N), PavedDrive (Y/N/P)

Feature Count After Encoding: 81 → 217 features

4.4 Text-Based Feature Representation

Objective: Combine descriptive categorical fields into composite text features that capture property characteristics holistically.

4.4.1 Composite Text Features (3)

1. property_location_type

- **Combination:** MSZoning + Neighborhood + Condition1

- **Purpose:** Captures WHERE and IN WHAT CONTEXT the property is located

- **Example:** "RL_CollgCr_Norm" = Residential Low Density in College Creek with Normal conditions

- **Unique Combinations:** 94

- **Rationale:** Location context is critical in real estate; combines zoning, area, and proximity factors

2. property_architecture

- **Combination:** BldgType + HouseStyle + RoofStyle

- **Purpose:** Captures ARCHITECTURAL DESIGN characteristics

- **Example:** "1Fam_2Story_Gable" = Single-family, two-story house with Gable roof

- **Unique Combinations:** 55

- **Rationale:** Architectural style affects buyer preferences and pricing

3. property_exterior

- **Combination:** Exterior1st + Exterior2nd + Foundation

- **Purpose:** Captures CONSTRUCTION MATERIALS and physical composition

- **Example:** "VinylSd_VinylSd_PConc" = Vinyl siding with poured concrete foundation

- **Unique Combinations:** 124

- **Rationale:** Material quality and foundation type impact durability and value

4.4.2 Encoding Strategy for Text Features

Method: Label Encoding

Rationale:

- High cardinality (55-124 unique combinations each)

- One-hot encoding would create 273 additional columns

- Label encoding preserves uniqueness while maintaining numeric format

- Appropriate for tree-based models and dimensionality reduction

Implementation:

```
 df['property_location_type'] = (
     df['MSZoning'] + '_' +
     df['Neighborhood'] + '_' +
     df['Condition1']
 )
 le = LabelEncoder()
 df['property_location_type'] = le.fit_transform(df['property_location_type'])
```

Feature Count After Text Features: 217 → 220 features

\newpage

5. Dimensionality Reduction

5.1 Motivation

Challenges with High-Dimensional Data:

- 220 features create risk of overfitting

- Multicollinearity detected (many correlated features)

- Curse of dimensionality affects model performance

- Computational complexity increases

Solution: Principal Component Analysis (PCA)

5.2 Feature Scaling

Method: StandardScaler (z-score normalization)

- **Formula:** $z = (x - \mu) / \sigma$

- **Result:** All features have mean=0, std=1

Rationale: PCA is sensitive to feature scales; standardization ensures equal contribution from all features.

5.3 Principal Component Analysis

Configuration:

- **Variance Threshold:** 95% (retain 95% of total variance)
- **Result:** 220 features → 136 principal components

Variance Explained:

- PC1: 11.2% of variance
- PC1-PC10: 45.3% cumulative
- PC1-PC50: 82.7% cumulative
- PC1-PC136: 95.0% cumulative (target achieved)

Benefits:

- Reduced dimensionality by 38.2%
- Eliminated multicollinearity (PCs are orthogonal)
- Retained 95% of information
- Improved computational efficiency

5.4 Student Random Feature Analysis in PCA

Question: Did the student_random_feature load significantly on any principal component?

Analysis Performed:

- Extracted loadings (weights) of student_random_feature across all 136 PCs
- Identified top 5 PCs with highest absolute loadings

Results:

- **Maximum Loading:** <0.05 (extremely weak)
- **Top PC Loading:** PC84 with loading of 0.043
- **Mean Absolute Loading:** 0.018 (near zero)

Conclusion: NO, the student_random_feature did NOT load significantly on any principal component.

Explanation:

- PCA captures **structured variance** in the data
- Random features lack systematic patterns or correlations
- The feature behaves as **noise** relative to real housing characteristics

- This confirms the feature's random nature and validates PCA's ability to distinguish signal from noise

\newpage

6. Exploratory Data Analysis

6.1 Visualizations Created

6.1.1 Missing Value Visualization

- **Type:** Bar chart + Heatmap
- **Purpose:** Identify patterns and extent of missingness
- **Key Finding:** Missingness concentrated in specific feature groups (pool, garage, basement features)

6.1.2 Distribution Analysis

- **Type:** Histograms with KDE overlay
- **Features:** SalePrice (before and after log transformation)
- **Key Finding:** Log transformation successfully normalized the right-skewed target variable (skewness: 1.88 → 0.12)

6.1.3 Correlation Heatmap

- **Type:** Heatmap with color gradient
- **Scope:** All 38 numeric features + student_random_feature
- **Key Findings:**
- **Top 5 Correlations with SalePrice:**
    1. OverallQual: 0.79
    2. GrLivArea: 0.71
    3. GarageCars: 0.64

    4. GarageArea: 0.62

    5. TotalBsmtSF: 0.61

- **student_random_feature correlation with SalePrice:** -0.006 (near zero, as expected)

6.1.4 Categorical vs. SalePrice Boxplots

- **Type:** Boxplots showing price distribution by category
- **Features:** OverallQual, ExterQual, KitchenQual, BsmtQual, Neighborhood, GarageType, Foundation, HeatingQC
- **Key Finding:** Clear price gradients across quality levels and neighborhoods

6.1.5 Scatterplots (Numeric vs. SalePrice)

- **Type:** Scatterplots with regression lines
- **Features:** GrLivArea, TotalBsmtSF, GarageArea, YearBuilt, 1stFlrSF, student_random_feature
- **Key Findings:**
- Strong linear relationships for area features
- Positive trend with YearBuilt
- **student_random_feature:** No discernible pattern (random scatter), confirming its random nature

6.2 Assignment Question 1: Student Random Feature Correlation

Question: Which 3 features appear most correlated with your random feature? Why does this occur?

Answer:

Top 3 Correlations:
1. **LowQualFinSF** (correlation: 0.046)
2. **KitchenAbvGr** (correlation: 0.036)
3. **MSSubClass** (correlation: 0.034)

Analysis:

These correlations are **extremely weak** (all < 0.05), which is expected and correct because:

1. **Random Nature:** The student_random_feature was generated using np.random.randint() with a specific seed, creating a uniform random distribution independent of actual housing data.

2. **Spurious Correlations:** These tiny correlations (0.03-0.05) represent random statistical noise, not meaningful relationships. In any dataset, some features will show small correlations by pure chance.

3. **Statistical Expectation:** With 38 numeric features, the expected number showing | correlation| > 0.03 by chance alone is approximately 2-3 features, matching our observation.

4. **Validation of Randomness:** The absence of strong correlations confirms the feature is truly random and not inadvertently capturing actual housing patterns.

Conclusion: The student_random_feature demonstrates no meaningful correlation with any real housing feature, validating its random generation and serving as a control feature for analysis.

\newpage

7. Final Datasets and Summary

7.1 Datasets Produced

7.1.1 processed_data_engineered.csv

- **Dimensions:** 1,459 rows × 218 columns
- **Content:**
- All original features (cleaned and transformed)
- 10 newly created numeric features
- 3 composite text-based features
- 176 one-hot encoded binary features
- Target: SalePrice_Log (log-transformed)

- **Use Case:** Ready for models that benefit from explicit features (linear models, interpretable models)

7.1.2 processed_data_pca.csv

- **Dimensions:** 1,459 rows × 137 columns

- **Content:**

- 136 principal components (PC1 through PC136)

- Target: SalePrice_Log

- **Variance Retained:** 95%

- **Use Case:** Ready for models sensitive to multicollinearity, high-dimensional models, faster training

7.2 Feature Engineering Pipeline Summary

```
 STAGE 0: Raw Data
├── Rows: 1,460
├── Columns: 81 (80 features + SalePrice)
├── Missing Values: 7,829 cells
└── Data Types: 43 categorical, 38 numeric


STAGE 1: Student Feature Addition
├── Generated student_random_feature (ID: 1100026)
└── Columns: 81 → 82


STAGE 2: Data Cleaning
├── Missing value treatment (contextual strategies)
├── Outlier removal (1 extreme case)
├── Missing Values: 7,829 → 0
└── Rows: 1,460 → 1,459


STAGE 3: Numeric Transformation
├── Log transformation applied to 29 skewed features
├── Target transformation: SalePrice → SalePrice_Log
└── Average skewness: 2.14 → 0.53


STAGE 4: Feature Creation
├── Created 10 new numeric features
└── Columns: 82 → 92


STAGE 5: Categorical Encoding
├── Ordinal encoding: 14 features
├── One-hot encoding: 29 features → 176 binary columns
├── Label encoding: 3 features
└── Columns: 92 → 217


STAGE 6: Text-Based Feature Representation
├── Created 3 composite text features
├── Label encoded each composite feature
└── Columns: 217 → 220


STAGE 7: Feature Scaling
├── StandardScaler applied to all 220 features
└── Result: Mean=0, Std=1 for all features


STAGE 8: Dimensionality Reduction (PCA)
```

```
├── Applied PCA with 95% variance threshold
├── Columns: 220 → 136 components
└── Variance retained: 95.0%


FINAL OUTPUT:
├── processed_data_engineered.csv (220 features)
└── processed_data_pca.csv (136 components)
```

\newpage

8. Key Decisions and Justifications

8.1 Missing Value Treatment

Decision: Context-based strategies rather than blanket imputation

Justification:

- Different features have different meanings for missingness

- "NA" in PoolQC means "no pool," not "unknown pool quality"

- Preserving semantic meaning improves model understanding

- Group-based imputation (LotFrontage by Neighborhood) respects spatial patterns

Evidence: Zero missing values achieved while maintaining data integrity

8.2 Feature Transformation

Decision: Log transformation for skewed features

Justification:

- Most ML algorithms assume approximately normal distributions

- Skewed features can disproportionately influence models

- Log transformation is theoretically sound for positive-valued features

- Commonly used in real estate pricing due to multiplicative price effects

Evidence: Skewness reduced from 2.14 to 0.53 (improvement of 75%)

8.3 Feature Creation

Decision: Create 10 domain-relevant numeric features + 3 text composites

Justification:
- Domain knowledge suggests total space matters more than individual rooms
- Age and renovation timing affect property values
- Binary indicators capture presence/absence of valuable features
- Composite text features capture how properties are actually described
- Interactions between categorical variables matter (location + context)

Evidence: Features align with real estate valuation principles

8.4 Encoding Strategy

Decision: Three-tier approach (ordinal, one-hot, label)

Justification:
- **Ordinal:** Preserves natural ordering in quality ratings
- **One-Hot:** Prevents false ordinality in nominal categories
- **Label Encoding:** Efficient for binary features and high-cardinality composites
- Different features require different treatments

Evidence: Encoding preserves information while maintaining model compatibility

8.5 Text-Based Features

Decision: Create 3 composite features from descriptive categoricals

Justification:
- Real estate descriptions are naturally composite ("suburban 2-story home")
- Captures feature interactions not visible in individual categories
- More semantically meaningful than separate encodings
- Label encoding chosen due to high cardinality (55-124 combinations)

Evidence: 273 unique combinations created across 3 features

8.6 Dimensionality Reduction

Decision: PCA with 95% variance retention

Justification:

- 220 features create overfitting risk

- High multicollinearity detected in correlation matrix

- PCA eliminates redundancy while preserving information

- 95% is standard threshold balancing information and dimensionality

- Computational efficiency improves with fewer features

Evidence: 38% reduction ($220 \rightarrow 136$) while retaining 95% variance

\newpage

9. Conclusion

9.1 Summary of Achievements

This feature engineering project successfully transformed a raw real estate dataset into two production-ready datasets through strategic application of data science principles:

Data Quality:

- Eliminated 7,829 missing values through contextual imputation strategies

- Removed 1 extreme outlier identified through IQR analysis

- Achieved 100% data completeness without compromising semantic meaning

Feature Engineering:

- Created 10 meaningful numeric features based on real estate domain knowledge

- Developed 3 composite text-based features capturing property descriptions

- Applied 29 log transformations to normalize skewed distributions

- Reduced target variable skewness by 94% ($1.88 \rightarrow 0.12$)

Encoding & Representation:

- Implemented three-tier encoding strategy (ordinal, one-hot, label)

- Expanded from 81 to 220 features through appropriate encoding
- Maintained semantic relationships and prevented false ordinality

Dimensionality Reduction:

- Applied PCA to reduce from 220 to 136 features (38% reduction)
- Retained 95% of original variance
- Eliminated multicollinearity through orthogonal components

Student Feature Integration:

- Generated and integrated student_random_feature (ID: 1100026)
- Analyzed correlation with all numeric features (max: 0.046, effectively zero)
- Evaluated PCA loadings (max: 0.043, no significant loading)
- Validated randomness through absence of meaningful patterns

9.2 Assignment Questions - Final Answers

Question 1: Which 3 features appear most correlated with your random feature? Why does this occur?

Answer: LowQualFinSF (0.046), KitchenAbvGr (0.036), MSSubClass (0.034). These extremely weak correlations represent random statistical noise, not meaningful relationships. The absence of strong correlations validates the feature's random generation and demonstrates that random features don't capture real housing patterns.

Question 2: After dimensionality reduction, did your random feature load significantly on any principal component?

Answer: No. Maximum loading was <0.05 across all 136 components. PCA captures structured variance, and random features lack systematic patterns. This confirms the feature behaves as noise relative to real housing characteristics and validates PCA's ability to distinguish signal from noise.

9.3 Key Insights

1. **Context Matters:** Missing values require contextual interpretation, not blanket strategies.

2. **Domain Knowledge is Critical:** Real estate pricing principles guided feature creation (TotalSF, HouseAge, etc.).

3. **Multiple Strategies Needed:** Different feature types require different encoding approaches.

4. **Text Features Add Value:** Composite categorical features capture interactions missed by individual encodings.

5. **Dimensionality Reduction is Essential:** PCA eliminates redundancy while preserving information.

6. **Random Features Don't Survive:** Student random feature showed no meaningful patterns, validating both its randomness and the effectiveness of correlation/PCA analysis.

9.4 Practical Applications

The resulting datasets are ready for:
- **Linear Models:** processed_data_engineered.csv with explicit features
- **Tree-Based Models:** Both datasets (trees handle multicollinearity)
- **Neural Networks:** processed_data_pca.csv for faster training
- **Ridge/Lasso Regression:** PCA components eliminate multicollinearity

9.5 Final Reflection

This assignment demonstrated that feature engineering is both an art and a science. Strategic decisions, grounded in domain knowledge and statistical principles, transform raw data into valuable model inputs. Every choice—from missing value imputation to dimensionality reduction—was made with clear justification and validated through evidence.

The integration of text-based feature representation added a crucial dimension, capturing how properties are naturally described in real estate contexts. This holistic approach to feature engineering exemplifies best practices in preparing data for machine learning applications.

Appendix A: Code Availability

Complete implementation available in: `FE_assignment.ipynb`

The Jupyter notebook contains:

- All code with detailed comments

- Comprehensive markdown explanations

- Visualizations with interpretations

- Output from all cells showing results

Appendix B: Data Dictionary

Original Features (Selected Key Features)

Feature	Type	Description
MSSubClass	Numeric	Building class
MSZoning	Categorical	General zoning classification
LotFrontage	Numeric	Linear feet of street connected to property
LotArea	Numeric	Lot size in square feet
OverallQual	Numeric	Overall material and finish quality (1-10)
YearBuilt	Numeric	Original construction year
TotalBsmtSF	Numeric	Total basement area in square feet
GrLivArea	Numeric	Above grade living area square feet
FullBath	Numeric	Full bathrooms above grade
BedroomAbvGr	Numeric	Bedrooms above grade
GarageArea	Numeric	Garage size in square feet
SalePrice	Numeric	Sale price in dollars (TARGET)

Engineered Features (New)

Feature	Type	Formula/Description
student_random_feature	Numeric	Random integers (1-100) + 4, seed=26
TotalSF	Numeric	TotalBsmtSF + 1stFlrSF + 2ndFlrSF
TotalBath	Numeric	FullBath + 0.5×HalfBath
HouseAge	Numeric	YrSold - YearBuilt
RemodAge	Numeric	YrSold - YearRemodAdd
TotalPorchSF	Numeric	Sum of all porch areas
HasPool	Binary	1 if PoolArea > 0, else 0
HasGarage	Binary	1 if GarageArea > 0, else 0
Has2ndFloor	Binary	1 if 2ndFlrSF > 0, else 0
HasBasement	Binary	1 if TotalBsmtSF > 0, else 0
HasFireplace	Binary	1 if Fireplaces > 0, else 0
property_location_type	Numeric (encoded)	MSZoning + Neighborhood + Condition1
property_architecture	Numeric (encoded)	BldgType + HouseStyle + RoofStyle
property_exterior	Numeric (encoded)	Exterior1st + Exterior2nd + Foundation

Appendix C: References and Tools

Software and Libraries

- **Python:** 3.11
- **pandas:** 2.3.3 - Data manipulation
- **NumPy:** 2.3.4 - Numerical computing
- **scikit-learn:** 1.7.2 - Machine learning tools (PCA, StandardScaler, LabelEncoder)
- **Matplotlib:** 3.10.7 - Visualization
- **Seaborn:** 0.13.2 - Statistical visualization
- **SciPy:** 1.16.3 - Scientific computing (stats)
- **Jupyter:** 1.1.1 - Notebook environment

Techniques Applied

- Missing value imputation (contextual, mode, median, zero-fill)
- Outlier detection (IQR method)
- Log transformation (log1p)
- Ordinal encoding
- One-hot encoding
- Label encoding
- Composite feature creation
- Standardization (z-score)
- Principal Component Analysis (PCA)

End of Report

Student Declaration:

I, Anik Das (Student ID: 2025EM1100026), declare that this report represents my original work and understanding. All code, analysis, and decisions documented herein were made independently, with justifications based on data science principles and domain knowledge.

Date: November 5, 2025
Signature: *Anik Das*