Feature Engineering Assignment ## House Price Prediction Dataset

Submitted By **Anik Das** **Student ID: 2025EM1100026**

Submitted To **BITS Pilani Digital** Course: Feature Engineering

Date of Submission November 4, 2025

Table of Contents

1. Introduction

This report documents my work on the feature engineering assignment using the Ames Housing dataset. The goal was to clean the data, create new features, handle categorical variables, and prepare the dataset for machine learning.

Dataset Overview

- **Rows:** 1,460 houses
- **Features:** 81 (80 features + SalePrice target)
- **Missing Values:** 7,829 cells across 19 columns
- **Target:** SalePrice (house prices)

Student Random Feature

As per the assignment requirements, I generated a random feature using my student ID: - Student ID: 2025EM1100026 - Last 7 digits: 1100026 - Seed: 26 (1100026 % 1000) - Offset: 4 (1100026 % 7) - Feature: `student_random_feature` with values from 5 to 103

2. Data Cleaning

2.1 Missing Value Treatment

I handled missing values based on what they actually mean in the context of housing data.

For features where "missing" means "doesn't exist": - Pool quality, garage features, basement features, etc. - Filled with 'None' for categorical or 0 for numeric - Makes sense because NA in PoolQC means no pool, not unknown quality

For LotFrontage: - Used median by neighborhood - Reasoning: Lot sizes vary by area, so neighborhood-specific imputation is better

For other truly missing values: - Used mode (most common value) for categorical - Only 1-2 cases, so simple approach works

Result: Zero missing values after treatment

2.2 Outliers

Used IQR method to detect outliers. Found 2 extreme cases with very large living area but unusually low prices - likely data errors or unusual sales. Removed them.

Final dataset: 1,459 rows (removed 1 outlier)

---

3. Feature Engineering

3.1 Transforming Skewed Features

Many features were right-skewed (most values low, few very high). Applied log transformation to 29 features including: - Area features (LotArea, GrLivArea, TotalBsmtSF, etc.) - The target variable (SalePrice)

Why log transform? - Makes distributions more normal - Helps models perform better - Common for real estate data where prices are multiplicative

Result: Target skewness reduced from 1.88 to 0.12

3.2 Creating New Features

Created 10 new features based on domain knowledge:

Aggregate features: 1. **TotalSF** = Basement + 1st floor + 2nd floor areas (total living space) 2. **TotalBath** = Full baths + 0.5 × half baths 3. **TotalPorchSF** = Sum of all porch areas

Temporal features: 4. **HouseAge** = Year sold - year built 5. **RemodAge** = Year sold - year remodeled

Binary indicators: 6-10. HasPool, HasGarage, Has2ndFloor, HasBasement, HasFireplace (1 if present, 0 if not)

Reasoning: Total space matters more than individual rooms. Age is more intuitive than year built. Binary flags capture presence of important features.

3.3 Text-Based Composite Features

Created 3 composite features by combining related categorical variables:

1. **property_location_type** = MSZoning + Neighborhood + Condition1
2. Example: "RL_CollgCr_Norm" = Residential area in College Creek with normal conditions
3. **property_architecture** = BldgType + HouseStyle + RoofStyle

4. Example: "1Fam_2Story_Gable" = Single-family two-story with gable roof

5. **property_exterior** = Exterior1st + Exterior2nd + Foundation

6. Example: "VinylSd_VinylSd_PConc" = Vinyl siding with concrete foundation

Why? Properties are described as combinations, not individual attributes. This captures interactions between features.

Encoding: Used label encoding because each has 50-120 unique combinations (too many for one-hot encoding).

---

4. Categorical Encoding

Used different strategies based on feature type:

4.1 Ordinal Encoding

For quality/condition features with natural order: - Quality scale: None(0) < Poor(1) < Fair(2) < Typical(3) < Good(4) < Excellent(5) - Applied to: ExterQual, KitchenQual, BsmtQual, HeatingQC, etc. (14 features)

4.2 One-Hot Encoding

For nominal features with no order: - Neighborhood, Foundation type, HouseStyle, etc. - Created 176 binary columns (24 original features)

4.3 Label Encoding

For binary features and the 3 text composites: - Street (Paved/Gravel), CentralAir (Y/N) - The 3 composite text features

Final feature count: 81 → 220 features

---

5. Dimensionality Reduction (PCA)

With 220 features, there's risk of overfitting and many features are correlated. Applied PCA to reduce dimensions while keeping most information.

Configuration: - Variance retention: 95% - Result: 220 features → 136 principal components

Benefits: - Reduced dimensionality by 38% - Eliminated multicollinearity (PCA components are uncorrelated) - Kept 95% of information - Faster model training

Student Random Feature in PCA

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

Answer: No, it didn't. - Maximum loading: 0.043 (PC84) - Mean absolute loading: 0.018

Why? Because it's random. PCA finds patterns in data, and random features don't have patterns. This confirms the feature is truly random.

---

6. Exploratory Analysis Results

6.1 Correlation with SalePrice

Top 5 correlated features: 1. OverallQual: 0.79 2. GrLivArea: 0.71 3. GarageCars: 0.64 4. GarageArea: 0.62 5. TotalBsmtSF: 0.61

student_random_feature correlation: -0.006 (essentially zero, as expected)

6.2 Student Random Feature Correlations

Assignment Question: Which 3 features are most correlated with your random feature?

Answer: 1. LowQualFinSF: 0.046 2. KitchenAbvGr: 0.036 3. MSSubClass: 0.034

These are extremely weak correlations (<0.05), which is correct. They represent random statistical noise, not real relationships. With 38 numeric features, you'd expect 2-3 to show tiny correlations just by chance. This validates the feature is truly random.

---

7. Final Datasets

7.1 processed_data_engineered.csv

- **Size:** 1,459 rows × 220 columns
- **Contents:** All engineered features, one-hot encoded variables
- **Use:** For models that benefit from explicit features (linear regression, interpretable models)

7.2 processed_data_pca.csv

- **Size:** 1,459 rows × 137 columns (136 PCA components + target)
- **Contents:** Principal components retaining 95% variance
- **Use:** For models sensitive to multicollinearity, faster training

Both datasets have: - Zero missing values - Log-transformed target - Standardized features (mean=0, std=1)

---

8. Summary

What was accomplished:

Data Cleaning: - Handled 7,829 missing values using context-based strategies - Removed 2 outliers - Achieved 100% data completeness

Feature Engineering: - Transformed 29 skewed features using log transformation - Created 10 new numeric features - Developed 3 composite text features - Reduced target skewness from 1.88 to 0.12

Encoding: - Ordinal encoding: 14 features - One-hot encoding: 24 features → 176 columns - Label encoding: 3 binary + 3 text features - Total: 81 → 220 features

Dimensionality Reduction: - Applied PCA: 220 → 136 components (38% reduction) - Retained 95% of variance - Eliminated multicollinearity

Student Feature Analysis: - Generated student_random_feature (ID: 1100026) - Correlation with SalePrice: -0.006 (negligible) - PCA loading: max 0.043 (no significant loading) - Confirms random nature and validates analysis methods

---

Key Learnings

1. **Context matters** - Missing values need different strategies based on what they mean
2. **Domain knowledge is important** - Real estate principles guided feature creation
3. **Different encodings for different features** - No one-size-fits-all approach
4. **Dimensionality reduction helps** - PCA reduces complexity while keeping information
5. **Random features don't survive analysis** - Validates that our methods work correctly

---

Files Submitted

1. **FE_assignment.ipynb** - Complete Jupyter notebook with code and analysis
2. **processed_data_engineered.csv** - Dataset with 220 engineered features
3. **processed_data_pca.csv** - Dataset with 136 PCA components
4. **This report** - Summary of work done

Student Declaration I confirm that this work is my own. All analysis and decisions were made based on the assignment requirements and my understanding of feature engineering principles. **Anik Das** Student ID: 2025EM1100026 Date: November 4, 2025