Assignment: Feature Engineering for Predictive Modeling - House Prices Dataset

Dataset

Kaggle Competition: House Prices - Advanced Regression Techniques

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Your ability will be judged by how **thoughtful and justified** your feature engineering decisions are.

Assignment Tasks

You are provided with a dataset containing property details and sale prices.
Your task is to **design and execute a feature engineering strategy** that transforms this raw dataset into a version ready for predictive modeling.

There are no fixed instructions - you must decide:

- Which data issues to clean, modify, or ignore
- Which transformations to apply and why
- Which features to create, merge, or remove
- Whether to apply dimensionality reduction, and to what extent
- How to represent text-based or categorical data meaningfully

Your final dataset should be suitable for machine learning, but you are not required to train or interpret a model.
The goal is to demonstrate **strategic thinking, technical correctness, and justification** of each choice you make during data preparation.

Your notebook should clearly show:

- The **decisions you made**, not just the operations you ran
- **Why** each technique was chosen
- Evidence (plots, metrics, reasoning) that supports those choices

Specific Instructions

- Each student must generate a **unique feature** based on their **student ID**.

- Let ID_last7 = last 7 digits of your student ID.
  (Example: for ID 2025EB1100221, ID_last7 = 1100221)

- Use the following random function code snippet to generate a new column in your dataset:

```
import numpy as np

def generate_student_feature(df, ID_last7):

      np.random.seed(ID_last7 % 1000)

      return np.random.randint(low=1, high=100, size=len(df)) +
(ID_last7 % 7)
```

- Add this new column to your dataset as: student_random_feature
- Treat it like any other numeric variable:
  - Include it in your **EDA**, **correlation analysis** and **dimensionality reduction** (if applied).
  - Decide how to scale or transform it, and justify your choice.

Exploratory Data Analysis (EDA)

To ensure originality and understanding:

- Include at least the following **visualisations** with short explanations:
  1. **Distribution plots** for key numeric features before and after transformation
  2. **Missing-value visualization** (bar chart or heatmap)
  3. **Correlation heatmap** for numeric features
  4. **Boxplots** showing categorical vs. SalePrice relationships
  5. **Scatterplots** showing engineered numeric features vs. SalePrice
- Visuals must include **your random feature**.
- Answer the following:
  1. Which 3 features appear most correlated with your *random feature*? Why do you think this occurs?
  2. After dimensionality reduction, did your random feature load significantly on any principal component? Explain briefly.

Deliverables

- .ipynb file must include both **code and output** (fully executed).
- Each key decision (cleaning, transformation, encoding, etc.) should have:
  - A **one-line explanation** before execution
  - **Evidence** (summary stats, plots, or metrics) after execution
  - Organise work with clear Markdown headers and concise comments
- **Report (small 1-2 pages PDF):** Summarising your feature engineering pipeline and reasoning behind each key decision.

Evaluation Rubric (Total = 40 Marks)

Criteria	Description	Marks
Data Familiarity & Initial Understanding	Demonstrates understanding of variable types, data distribution, and relationships; identifies potential data issues independently.	4
Data Cleaning Decisions	Handles missing values, outliers, or inconsistencies logically with clear justification and before–after evidence.	6
Numeric Feature Engineering	Applies appropriate scaling, transformation, or discretization techniques thoughtfully; avoids mechanical use.	6
Feature Creation & Encoding	Constructs meaningful new variables (including random-function feature); applies correct encoding for categorical data.	8
Dimensionality Reduction & Correlation Handling	Identifies redundancy and multicollinearity; applies or justifies PCA or similar technique appropriately.	6
Text-Based Feature Representation	Combines descriptive fields into text; cleans and encodes text meaningfully.	6
Documentation & Clarity	Notebook readability, logical flow, commentary quality, and clarity in the PDF report.	2
Bonus	Shows exceptional creativity or deeper exploration beyond syllabus expectations.	2

- You are not graded on model accuracy - the focus is *data reasoning*.
- Avoid using automated pipelines without explanation.
- Every transformation must have a clear "why."
- Simplicity with strong justification will score higher than complexity without reasoning.

Evaluation Criteria

Out of: **40 points**

Submission Details

Submission type
File submission
Files allowed per submission
Unlimited
Allowed files
Custom File Types (.ipynb,.pdf)
Number of submissions
Only the most recent submission is kept

Criterion	Description
Data Audit & Availability Check	Furnish lines of code on inspecting dataset shape, data types, counts, missing values%, duplicates, unique value counts, other relevant, and confirm availability of columns relevant to the task, etc.
Exploratory Data Analysis (EDA)	Furnish lines of code and provide relevant summary statistics, distributions, and visualizations (histograms, boxplots, scatterplots, time-series plots where applicable, correlation tables) and derive insights/hypotheses tied to the problem statement and any other relevant. When specific types of plots not applicable include one-line Justification in bold directly under relevant notebook section as Markdown cell.
Data Cleaning	Handle missing values (imputation/removal), duplicates, incorrect data types, outliers (with justification), irrelevant columns removal, and any dataset-specific fixes. When dataset requires no cleaning, include a one-line Justification (Not Applicable).
Feature Creation / Transformation	Furnish lines of code to create or transform features appropriate to problem statement and dataset (e.g., lag features, RFM metrics, interaction terms, encodings). Ensure each feature is relevant, justified and useful. If few features are needed, include Justification in bold under the relevant notebook section as Markdown.
Feature Selection / Dimensionality Reduction	Furnish lines of code to apply methods like filter, wrapper, embedded methods, feature-importance, correlation filtering, RFE, PCA, etc. Include brief Justification in bold under notebook section or if not done, justify why.
Feature Evaluation / Quick Checks	Furnish code for simple checks proving engineered features are meaningful (correlation with target, feature-importance, baseline tests). For unsupervised tasks, justify logically or with data or if not applicable, justify with few lines in bold under notebook section.