

**Lahore College for Women University, Lahore**

**Department: Software Engineering**

**Submitted By: Iram Ahmad**

**Roll no: 2225165111**

**Submitted to: Sir Mohsin Nazir**

**Semester: VII**

**Session: 2022-2026**

**Assignment no 4**

**Course: Applied Data Science with AI**

**Semester:** BSSE 7th  
**Week #:** 4  
**Student Name:** Iram Ahmad  
**Roll Number:** 2225165111  
**Project Title:** House Price Prediction  
**GitHub Link:** https://github.com/Iram-Ahmad/Data-Science-AI-Course

**1. Reading Summary (½–1 page)**

**Reading Material for this Week:**

* Khan Academy — Statistics & Probability (Basics: mean, median, mode, variance).
* “Stats for Data Science” notes (course handout / summary materials).
* Pandas & NumPy quick reference for statistical functions.

**Key Learnings:**

* **Mean / Median / Mode:** Central tendency measures that summarize a dataset’s center. Median is robust to outliers.
* **Variance & Standard Deviation:** Measures of spread; variance shows how much values disperse from the mean.
* **Correlation:** Pearson correlation measures linear relationships between numeric variables and helps identify features related to the target variable (SalePrice).

**Reflection:**  
Understanding descriptive statistics and correlation is critical for feature selection in the House Price Prediction project. These measures help identify which variables have the most direct linear association with price and guide further feature engineering and modeling.

**2. Classroom Task Documentation**

**Task Performed in Class:**

* Calculated **mean, median, mode, variance** for key numeric features (e.g., SalePrice, GrLivArea, OverallQual).
* Created a **correlation matrix** and plotted a heatmap to visually inspect relationships among numeric features.
* Encoded low-cardinality categorical features (label encoding) to check their correlation with SalePrice.

**Screenshots / Code Snippets:**

# Example: mean, median, variance for SalePrice

df['SalePrice'].mean(), df['SalePrice'].median(), df['SalePrice'].var()

# Correlation and top features

num\_df = df.select\_dtypes(include=['number'])

corr = num\_df.corr()

corr['SalePrice'].abs().sort\_values(ascending=False).head(10)

**3. Weekly Assignment Submission**

**Assignment Title:** Correlation Analysis & Feature Selection

**Steps Taken:**

1. Loaded cleaned dataset (cleaned\_house\_prices.csv).
2. Computed basic statistics (mean, median, mode, variance) for numeric features.
3. Calculated Pearson correlation for numeric features and encoded low-cardinality categorical variables for inspection.
4. Produced a ranked list of features most strongly correlated with SalePrice and saved results to week4\_top\_correlations.csv.

**Output:**

* week4\_top\_correlations.csv (uploaded to repo) — lists numeric features ranked by absolute correlation with SalePrice.
* The top 3 features (example — replace after running): **OverallQual, GrLivArea, GarageCars**.

**Note:** Run the notebook to get exact features for your dataset; the example above matches common results for the Ames dataset.

**Challenges Faced:**

* Need to treat categorical variables carefully — Pearson correlation applies only to numeric data.
* Deciding whether to encode categorical features or to use ANOVA/other tests for categorical vs numeric relationships.

**GitHub Link:**  
https://github.com/Iram-Ahmad/Data-Science-AI-Course

**4. Project Progress Milestone**

**This Week’s Milestone:**  
Identified candidate predictive variables via correlation analysis; prepared Week4\_Statistics\_and\_Correlation.ipynb.

**Next Week’s Goal:**  
Start **feature engineering** and prepare features for model building (transformations, encoding, scaling).

**5. Self-Evaluation (Check one)**

☑ I completed all tasks on time.  
⬜ I partially completed the tasks.  
⬜ I struggled with this week’s tasks and need help.

**6. Questions for Instructor**

* Should we drop features with only weak correlation, or keep them and let model regularization decide?
* For categorical variables with many categories, do you prefer one-hot encoding or target encoding in this course?