

**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 2**

**Course: Applied Data Science with AI**

**Semester:** BSSE 7th  
**Week #:** 2  
**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:**

* Pandas Documentation (<https://pandas.pydata.org/docs/>)
* NumPy Documentation (<https://numpy.org/doc/>)

**Key Learnings:**

1. Learned how to use **Pandas** for data manipulation — including reading CSV files, exploring data frames, and handling missing data.
2. Understood how **NumPy** supports numerical operations and serves as the foundation for Pandas — especially when dealing with arrays and mathematical computations.
3. Explored techniques for **data cleaning**, such as detecting null values, removing duplicates, and replacing missing values with the mean or mode.

**Reflection:**  
These readings were very helpful for preparing my dataset for machine learning. I used the knowledge from Pandas to clean my **House Price Prediction dataset** — removing missing values and duplicates to make it ready for analysis. Understanding NumPy helped me handle numerical columns efficiently while performing calculations such as replacing missing numeric data with average values.

**2. 🧠 Classroom Task Documentation**

**Task Performed in Class:**

* Practiced **data cleaning techniques** using Pandas and NumPy.
* Removed duplicate rows, handled missing values, and detected outliers.
* Discussed the importance of data preprocessing before applying machine learning models.

**Screenshots / Code Snippets:**  
*(Sample code used in class)*

# Remove duplicates

df.drop\_duplicates(inplace=True)

# Handle missing values

df['LotFrontage'].fillna(df['LotFrontage'].mean(), inplace=True)

df['GarageType'].fillna('None', inplace=True)

# Check null values

df.isnull().sum().head(10)

**3. 💻 Weekly Assignment Submission**

**Assignment Title:** Data Cleaning and Preprocessing

**Steps Taken:**

1. Loaded the raw dataset (house\_prices.csv) from Kaggle.
2. Checked for duplicates, missing values, and outliers.
3. Cleaned the data by filling missing numerical values with mean and categorical ones with “None.”
4. Saved the cleaned dataset as cleaned\_house\_prices.csv.

**Output:**  
Before cleaning, several columns had missing values (e.g., *LotFrontage*, *GarageType*, *PoolQC*).  
After cleaning, most columns had no missing data, and duplicates were removed.

| **Column** | **Missing (Before)** | **Missing (After)** |
| --- | --- | --- |
| LotFrontage | 259 | 0 |
| GarageType | 81 | 0 |
| PoolQC | 1453 | 1453 (too few values → dropped later) |

**Challenges Faced:**

* Initially got a FileNotFoundError due to incorrect file path.
* The dataset contained many columns with missing values, which required careful handling.
* Faced confusion about uploading the cleaned dataset to GitHub (resolved after understanding folder structure).

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

**4. 🚀 Project Progress Milestone**

**This Week’s Milestone:**  
✅ Downloaded and loaded the dataset.  
✅ Performed data cleaning (handled missing values and duplicates).  
✅ Uploaded the cleaned dataset to GitHub.

**Next Week’s Goal:**  
➡ Perform **Exploratory Data Analysis (EDA)** to understand relationships between features and target variable (SalePrice).

**5. ✅ Self-Evaluation**

☑ **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 (Optional)**

* What’s the best way to visualize outliers before model training?