

**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 3**

**Course: Applied Data Science with AI**

**Semester:** BSSE 7th  
**Week #:** 3  
**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:**

* “Data Visualization with Python” – from *Python Data Science Handbook* by Jake VanderPlas
* Matplotlib and Seaborn official documentation
* “Storytelling with Data: A Data Visualization Guide for Business Professionals” – Chapter 2: Choosing the Right Chart

**● Key Learnings:**

* **Matplotlib** is the fundamental plotting library in Python for creating static, interactive, and publication-quality visualizations.
* **Seaborn** provides a higher-level interface for attractive statistical graphics and is great for exploring data distributions and relationships.
* Selecting the right type of visualization (e.g., histogram, boxplot, scatterplot, heatmap) depends on the nature of the data and what insights you want to highlight.

**● Reflection:**  
These readings helped me understand how to visualize the housing dataset effectively. Using visualizations like histograms and boxplots made it easier to identify outliers, trends, and relationships between house features (like OverallQual and SalePrice). This understanding directly supports building a better predictive model in the upcoming weeks.

**2. Classroom Task Documentation**

**● Task Performed in Class:**

* Loaded the cleaned house price dataset into a Pandas DataFrame.
* Created multiple visualizations using Matplotlib and Seaborn to understand data distribution and relationships:
  1. Histogram of SalePrice
  2. Scatter plot of GrLivArea vs SalePrice
  3. Boxplot of OverallQual vs SalePrice
  4. Correlation heatmap of numeric features

**● Screenshots / Code Snippets:**

sns.histplot(df['SalePrice'], bins=30, kde=True)

plt.title('Distribution of House Prices')

plt.show()

sns.scatterplot(x='GrLivArea', y='SalePrice', data=df)

plt.title('Living Area vs Sale Price')

plt.show()

sns.boxplot(x='OverallQual', y='SalePrice', data=df, palette='viridis')

plt.title('Overall Quality vs Sale Price')

plt.show()

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

sns.heatmap(corr, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

**3. Weekly Assignment Submission**

**● Assignment Title:** Data Visualization and Insights

**● Steps Taken:**

1. Imported and cleaned the house price dataset.
2. Created visualizations to analyze data distribution and relationships.
3. Interpreted insights from each plot and documented observations.

**● Output (Sample Results/Plots):**

* Most houses are priced between **100,000 and 200,000**.
* The price distribution is **right-skewed**, meaning a few houses are very expensive.
* A strong positive correlation exists between GrLivArea and SalePrice.
* OverallQual (overall material and finish quality) is one of the most influential factors in house prices.

**● Challenges Faced:**

* Encountered a FutureWarning with Seaborn’s palette parameter in boxplot (resolved by adjusting syntax).
* Needed to handle non-numeric columns for correlation analysis using select\_dtypes().

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

**4. Project Progress Milestone**

**● This Week’s Milestone:**  
Created and analyzed multiple visualizations to understand feature relationships in the dataset.

**● Next Week’s Goal:**  
Perform **feature correlation and selection**, and start preparing data for **model training** (regression analysis).

**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 (Optional)**

* How can we decide which features to keep or remove before applying regression models?
* What’s the best way to visualize interactions between more than two features?