**Lab Exercise 2– Plotting Your Data in SageMaker Jupyter Notebooks**

**1. Lab Title**

**Visuals: Plotting and Visualizing Data in Amazon SageMaker Jupyter Notebooks (with S3 Data Loading)**

**2. Objective**

The objective of this lab is to:

* Learn how to **load datasets directly from Amazon S3** into SageMaker Jupyter notebooks.
* Explore **data visualization techniques** using **Matplotlib** and **Seaborn**.
* Understand relationships, trends, and patterns in data through different types of plots.

By the end of this lab, you will be able to:  
Load data from S3 into a Pandas DataFrame  
Create multiple visualizations for analysis  
Interpret data trends from visuals

**3. Prerequisites**

Before starting, ensure that:

* You have an **AWS Account** with access to **Amazon SageMaker Studio** and **Amazon S3**.
* You have a **Python 3 (Data Science)** kernel running in SageMaker Studio.
* You have a dataset (e.g., house\_prices\_dataset.csv) stored in your **S3 bucket**.
* You have basic knowledge of **Python** and **data visualization libraries**.

**4. Dataset Overview**

We will visualize a **housing dataset** that contains property information for various cities.

| **Column** | **Description** |
| --- | --- |
| id | Unique identifier |
| location | City name |
| area | Property size in square feet |
| bedrooms | Number of bedrooms |
| bathrooms | Number of bathrooms |
| price | Property price (in INR) |

**5. Lab Setup**

**Step 1: Launch SageMaker Studio**

1. Log in to the **AWS Management Console**.
2. Navigate to **Amazon SageMaker → SageMaker Studio**.
3. Select your **user profile** and click **Launch app → Studio**.
4. Once the environment loads, open a **new notebook** and select the **Python 3 (Data Science)** kernel.

Rename the notebook:

Visualizing\_Data\_from\_S3.ipynb

**Step 2: Import Required Libraries**

In your first code cell, import the required libraries:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import boto3

# Enable inline plotting

%matplotlib inline

# Set Seaborn style

sns.set(style="whitegrid")

**Step 3: Load Data from Amazon S3**

Now we’ll use **boto3** — the AWS SDK for Python — to load the dataset directly from an S3 bucket.

# --- Configure your S3 bucket details ---

bucket\_name = "your-bucket-name" # Replace with your S3 bucket name

file\_key = "datasets/house\_prices\_dataset.csv" # S3 path to your dataset

# Initialize S3 client

s3 = boto3.client('s3')

# Generate S3 path for reading

s3\_path = f"s3://{bucket\_name}/{file\_key}"

# Load data into Pandas DataFrame

df = pd.read\_csv(s3\_path)

# Display first few rows

df.head()

**Expected Output:**

A preview of the first few rows of the dataset showing columns like location, area, and price.

**Step 4: Inspect the Dataset**

# Check dataset information

df.info()

# Summary statistics

df.describe()

**Interpretation:**  
This step confirms data structure, column types, and general statistics such as average price and area.

**6. Lab Tasks — Data Visualization**

**Task 1: Histogram — Distribution of House Prices**

plt.figure(figsize=(8,5))

sns.histplot(df['price'], bins=20, kde=True, color='royalblue')

plt.title("Distribution of House Prices")

plt.xlabel("Price (INR)")

plt.ylabel("Frequency")

plt.show()

**Purpose:** Understand how property prices are distributed (normal, skewed, etc.)

**Task 2: Scatter Plot — Area vs Price**

plt.figure(figsize=(8,5))

sns.scatterplot(x='area', y='price', data=df, hue='bedrooms', palette='coolwarm', s=100)

plt.title("Relationship Between Area and Price")

plt.xlabel("Area (sq ft)")

plt.ylabel("Price (INR)")

plt.show()

**Purpose:** Analyze how area and number of bedrooms influence property price.

**Task 3: Bar Chart — Average Price by Location**

avg\_price\_df = df.groupby('location')['price'].mean().reset\_index().sort\_values(by='price', ascending=False)

plt.figure(figsize=(10,6))

sns.barplot(x='location', y='price', data=avg\_price\_df, palette='viridis')

plt.title("Average House Price by City")

plt.xticks(rotation=45)

plt.xlabel("City")

plt.ylabel("Average Price (INR)")

plt.show()

**Purpose:** Compare average property prices across cities.

**Task 4: Box Plot — Price Distribution by Bedroom Count**

plt.figure(figsize=(8,5))

sns.boxplot(x='bedrooms', y='price', data=df, palette='Set2')

plt.title("House Price Distribution by Bedrooms")

plt.xlabel("Bedrooms")

plt.ylabel("Price (INR)")

plt.show()

**Purpose:** Identify variation and outliers for each bedroom category.

**Task 5: Heatmap — Correlation Between Variables**

plt.figure(figsize=(8,6))

# Select only numeric columns

numeric\_df = df.select\_dtypes(include=['float64', 'int64'])

# Compute correlation matrix

corr\_matrix = numeric\_df.corr()

# Plot the heatmap

sns.heatmap(corr\_matrix, annot=True, cmap="YlGnBu", fmt=".2f")

plt.title("Feature Correlation Heatmap (Numeric Columns Only)")

plt.show()

**Purpose:** Discover which features most strongly influence price (e.g., area, bedrooms).

**7. Save and Export Visualizations**

You can save generated plots for reports or presentations.

plt.figure(figsize=(8,5))

sns.histplot(df['price'], bins=30, kde=True, color='blue')

plt.title("Distribution of House Prices")

plt.savefig("price\_distribution.png")

plt.close()

print("Plot saved as price\_distribution.png")

Optionally upload it to S3:

s3.upload\_file("price\_distribution.png", bucket\_name, "results/price\_distribution.png")

**8. Validation & Observation**

After completing the lab:  
Dataset loaded successfully from S3  
All plots generated correctly  
Visuals show meaningful relationships (e.g., area vs price)  
Saved and optionally uploaded plots to S3

**9. Cleanup**

To avoid unnecessary costs:

* Save your notebook in SageMaker Studio.
* Shut down all notebook instances (**File → Shut down → All Kernels**).
* Delete temporary plots or large data files.