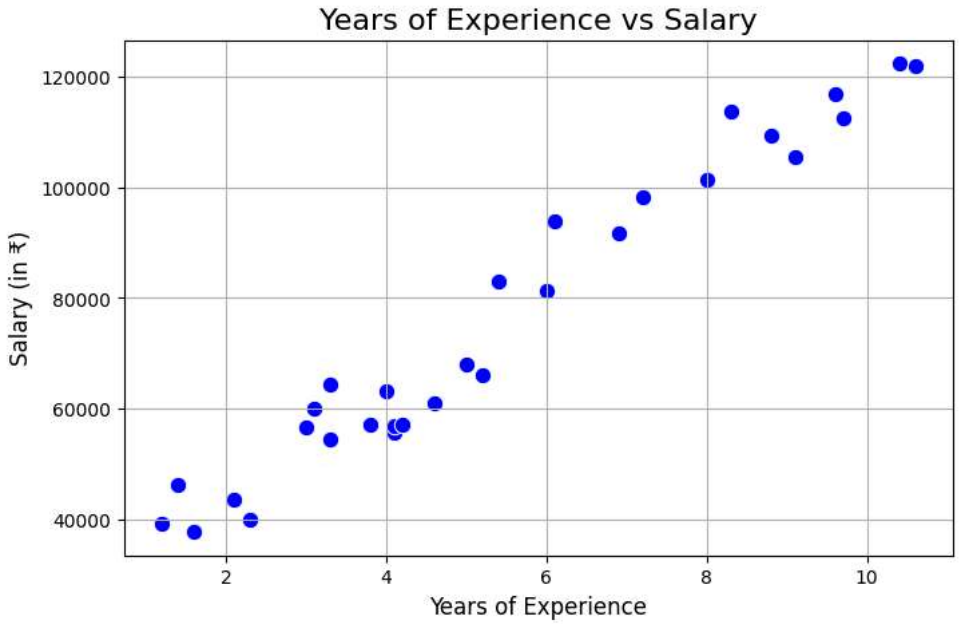
```python
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load the dataset (replace with the correct path if needed)
file_path = "/content/Salary_dataset.csv"
df = pd.read_csv(file_path)

# Step 1: EDA
# Remove unnecessary column
df_cleaned = df.drop(columns=["Unnamed: 0"])

# Display basic info and statistics
print("Dataset Overview:\n", df_cleaned.head())
print("\nSummary Statistics:\n", df_cleaned.describe())

# Scatter plot to visualize the relationship
plt.figure(figsize=(8, 5))
sns.scatterplot(data=df_cleaned, x="YearsExperience", y="Salary", color="blue", s=80)
plt.title("Years of Experience vs Salary", fontsize=16)
plt.xlabel("Years of Experience", fontsize=12)
plt.ylabel("Salary (in ₹)", fontsize=12)
plt.grid(True)
plt.show()

# Correlation calculation
correlation = df_cleaned.corr().loc["YearsExperience", "Salary"]
print(f"\nCorrelation between YearsExperience and Salary: {correlation:.2f}")

# Step 2: Prepare Data for Linear Regression
X = df_cleaned[["YearsExperience"]]  # Independent variable
y = df_cleaned["Salary"]  # Dependent variable

# Split the dataset into training and testing sets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 3: Train the Linear Regression Model
model = LinearRegression()
model.fit(X_train, y_train)

# Step 4: Make Predictions
y_pred = model.predict(X_test)

# Step 5: Evaluate the Model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"\nMean Squared Error (MSE): {mse:.2f}")
print(f"R-squared (R2 Score): {r2:.2f}")

# Step 6: Visualize the Regression Line
plt.figure(figsize=(8, 5))
sns.scatterplot(x=X_train["YearsExperience"], y=y_train, color="blue", label="Training Data")
sns.scatterplot(x=X_test["YearsExperience"], y=y_test, color="red", label="Test Data", marker="x")
plt.plot(X_test, y_pred, color="green", linewidth=2, label="Regression Line")
plt.title("Simple Linear Regression: Years of Experience vs Salary", fontsize=16)
plt.xlabel("Years of Experience", fontsize=12)
plt.ylabel("Salary (in ₹)", fontsize=12)
plt.legend()
plt.grid(True)
plt.show()

```

```
Dataset Overview:
    YearsExperience   Salary
0              1.2  39344.0
1              1.4  46206.0
2              1.6  37732.0
3              2.1  43526.0
4              2.3  39892.0

Summary Statistics:
       YearsExperience         Salary
count        30.000000      30.000000
mean          5.413333   76004.000000
std           2.837888   27414.429785
min           1.200000   37732.000000
25%           3.300000   56721.750000
50%           4.800000   65238.000000
75%           7.800000  100545.750000
max          10.600000  122392.000000
```




```
Correlation between YearsExperience and Salary: 0.98

Mean Squared Error (MSE): 49830096.86
R-squared (R2 Score): 0.90
```