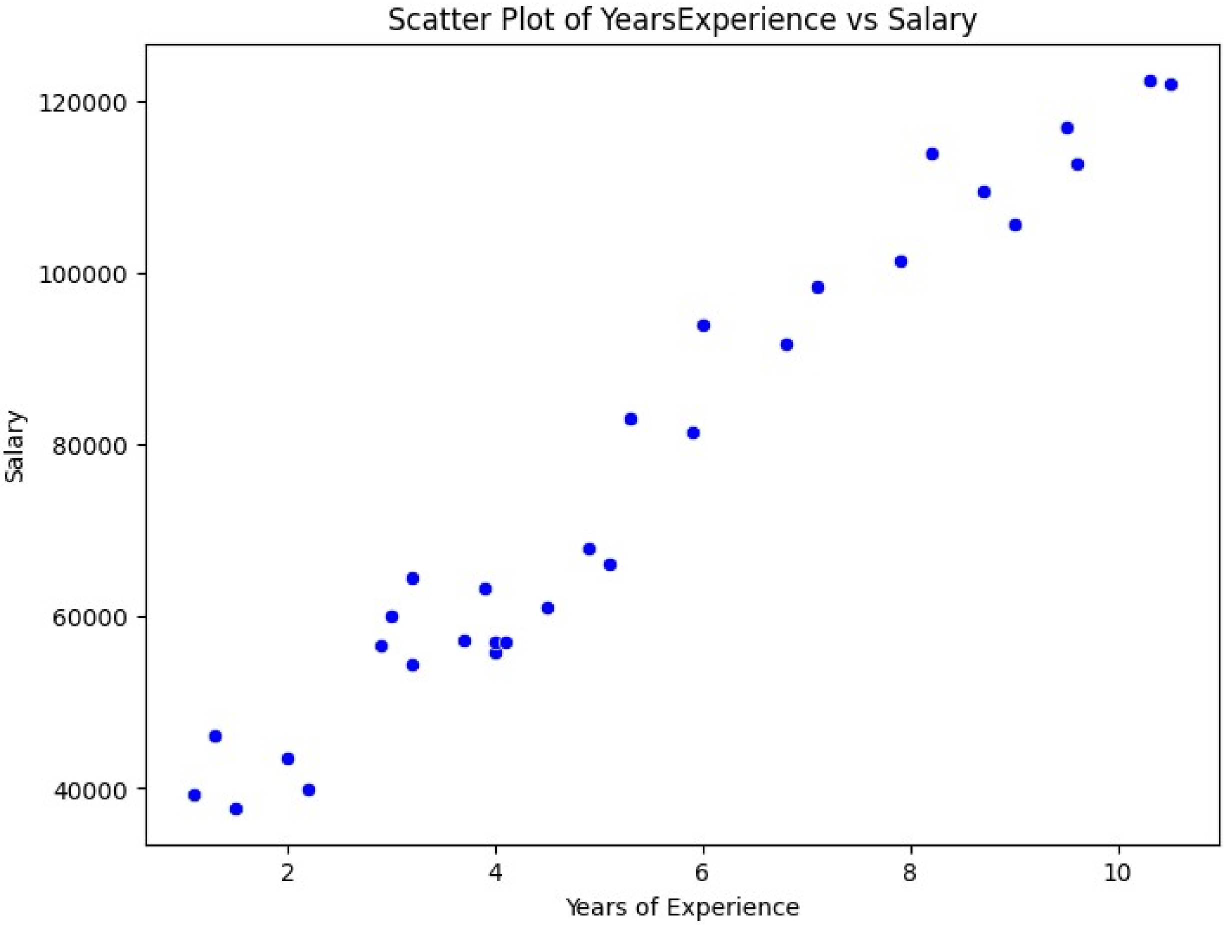
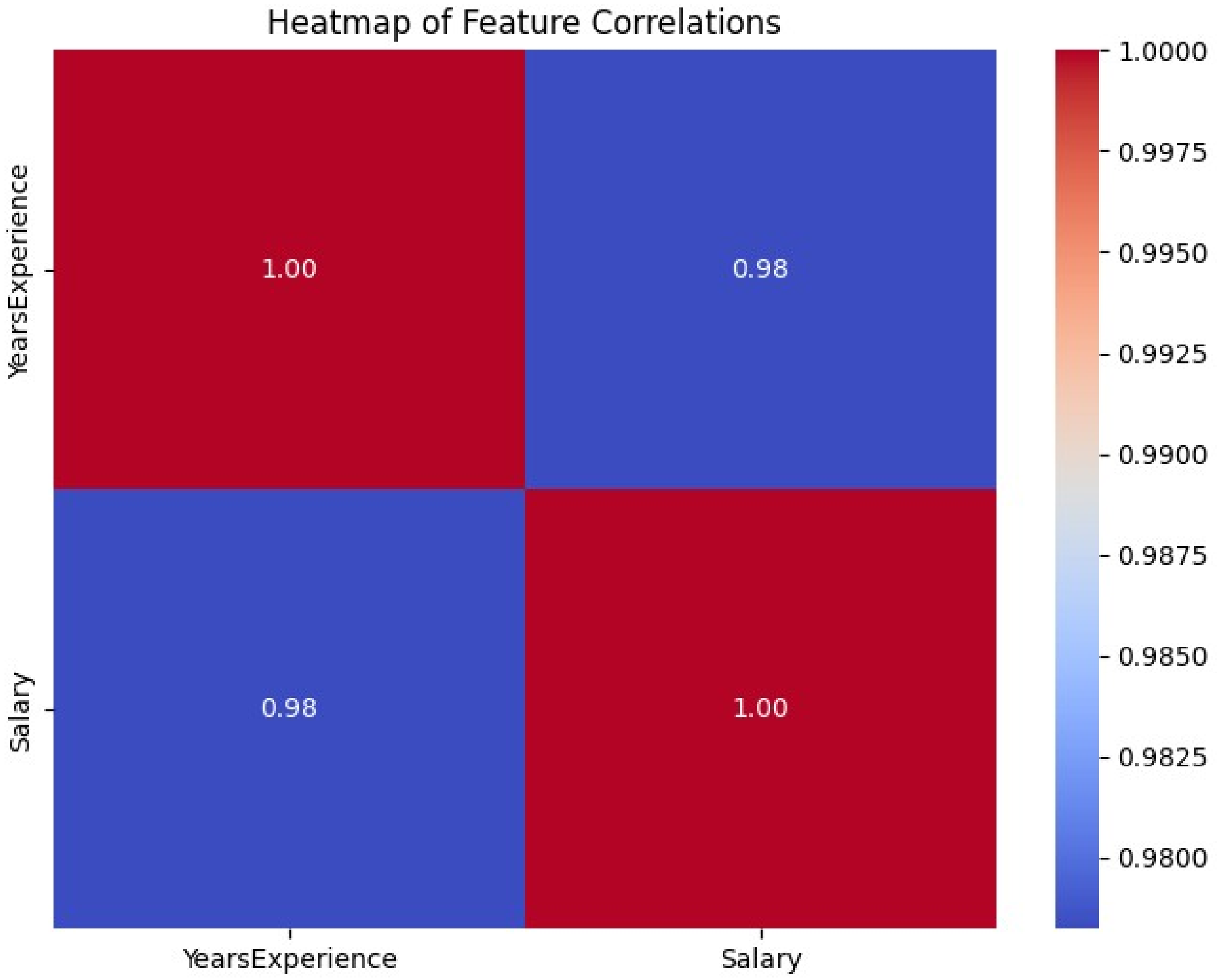
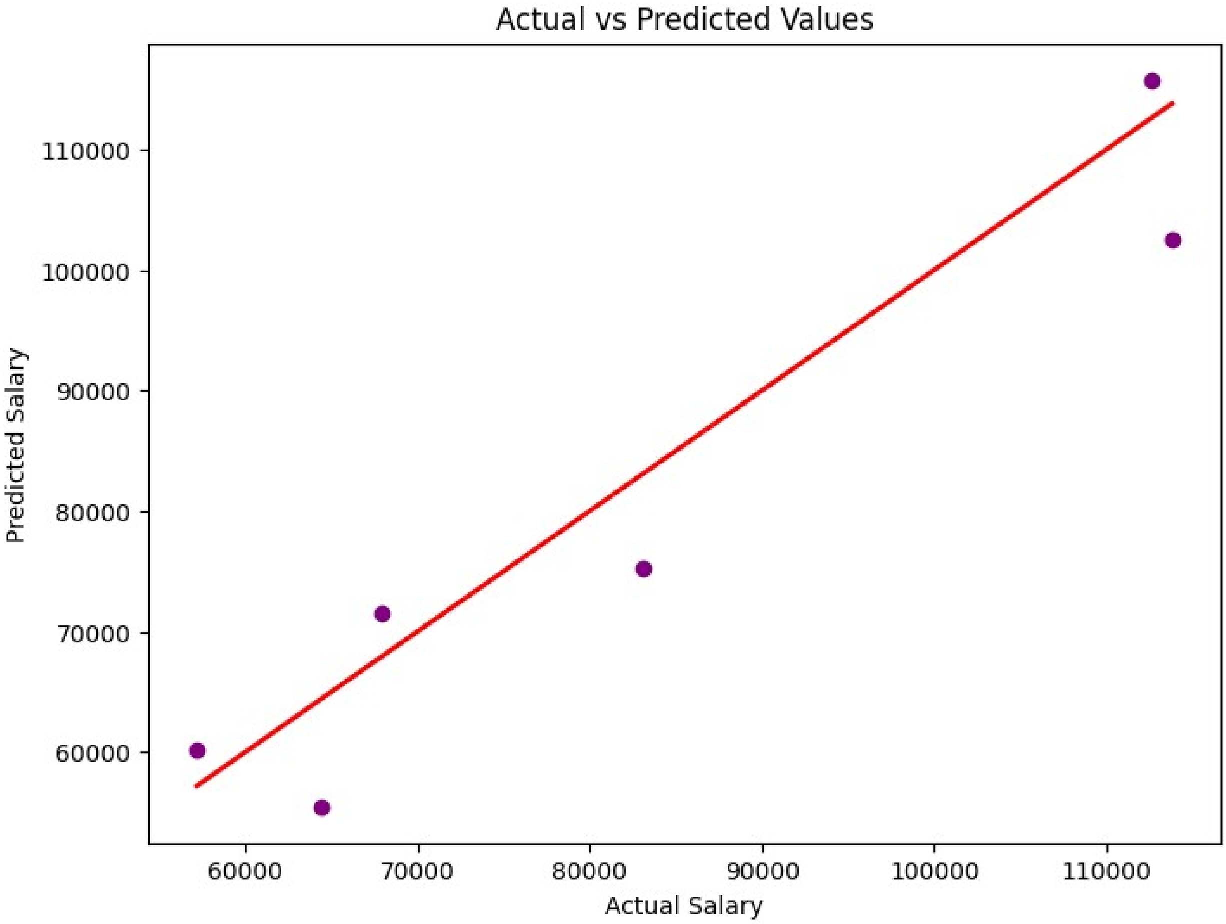
```python
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error,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

# Step 1: Dataset Import and Preprocessing
# Load the dataset
file_path = "Salary_Data.csv"
data = pd.read_csv(file_path)

# Display first few rows and dataset structure
print("Dataset Structure:")
print(data.info())
print("\nFirst few rows of the dataset:")
print(data.head())

# Check for missing values
print("\nMissing values in each column:")
print(data.isnull().sum())

# Basic statistics
print("\nDataset Statistics:")
print(data.describe())

# Step 2: Exploratory Data Analysis (EDA)
# Scatter plot to identify relationships
plt.figure(figsize=(8, 6))
sns.scatterplot(data=data, x="YearsExperience", y="Salary",
color='blue')
plt.title("Scatter Plot of YearsExperience vs Salary")
plt.xlabel("Years of Experience")
plt.ylabel("Salary")
plt.show()

# Heatmap to identify correlations
plt.figure(figsize=(8, 6))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Heatmap of Feature Correlations")
plt.show()

# Step 3: Linear Regression Model Implementation
# Split dataset into training and testing sets
X = data[["YearsExperience"]]  # Input feature(s)
y = data["Salary"]  # Target variable
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Create and train the model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Step 4: Evaluation Metrics
# Calculate performance metrics
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("\nModel Evaluation Metrics:")
print(f"Mean Absolute Error (MAE): {mae:.2f}")
print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"R-squared Value: {r2:.2f}")

# Step 5: Visualizing the Regression Line
# Plot actual vs predicted values
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, color='purple')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)],
color='red', linewidth=2)
plt.title("Actual vs Predicted Values")
plt.xlabel("Actual Salary")
plt.ylabel("Predicted Salary")
plt.show()

# Display regression line on scatter plot of training data
plt.figure(figsize=(8, 6))
plt.scatter(X_train, y_train, color='blue', label="Training Data")
plt.plot(X_train, model.predict(X_train), color='red', linewidth=2,
label="Regression Line")
plt.title("Regression Line on Training Data")
plt.xlabel("Years of Experience")
plt.ylabel("Salary")
plt.legend()
plt.show()

# Step 6: Conclusion
print("\nConclusion:")
print("The linear regression model has been trained and evaluated on
the dataset. "
      "The R-squared value indicates the proportion of variance in the
target variable "
      "explained by the input features. Improvements may include
adding more features or trying advanced models.")
```

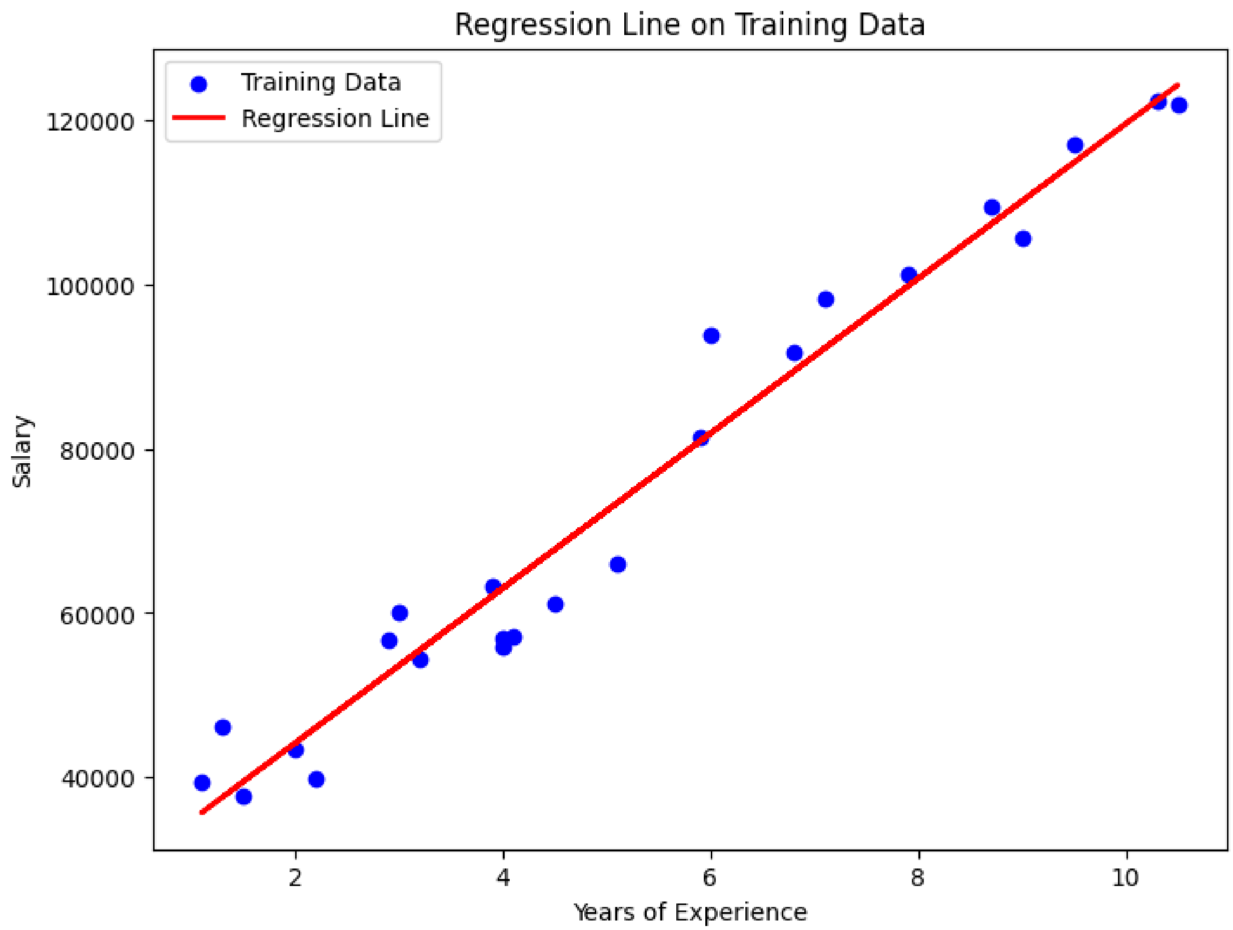
```
Dataset Structure:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   YearsExperience  30 non-null     float64
 1   Salary           30 non-null     float64
dtypes: float64(2)
memory usage: 608.0 bytes
None

First few rows of the dataset:
   YearsExperience   Salary
0              1.1  39343.0
1              1.3  46205.0
2              1.5  37731.0
3              2.0  43525.0
4              2.2  39891.0

Missing values in each column:
YearsExperience    0
Salary             0
dtype: int64

Dataset Statistics:
       YearsExperience         Salary
count        30.000000      30.000000
mean          5.313333   76003.000000
std           2.837888   27414.429785
min           1.100000   37731.000000
25%           3.200000   56720.750000
50%           4.700000   65237.000000
75%           7.700000  100544.750000
max          10.500000  122391.000000
```


Scatter Plot of YearsExperience vs Salary
Salary
120000
100000
80000
60000
40000
2
4
6
8
10
Years of Experience



```
Model Evaluation Metrics:
Mean Absolute Error (MAE): 6286.45
Mean Squared Error (MSE): 49830096.86
R-squared Value: 0.90
```


Actual vs Predicted Values
Predicted Salary
110000
100000
90000
80000
70000
60000
Actual Salary
60000
70000
80000
90000
100000
110000



```
Conclusion:
The linear regression model has been trained and evaluated on the
dataset. The R-squared value indicates the proportion of variance in
the target variable explained by the input features. Improvements may
include adding more features or trying advanced models.
```