Dataset Loading and Understanding

```
import pandas as pd
import numpy as np
df = pd.read_csv('/content/Salary_dataset.csv')
print(df)
```

```
    Unnamed: 0  YearsExperience    Salary
0            0              1.2   39344.0
1            1              1.4   46206.0
2            2              1.6   37732.0
3            3              2.1   43526.0
4            4              2.3   39892.0
5            5              3.0   56643.0
6            6              3.1   60151.0
7            7              3.3   54446.0
8            8              3.3   64446.0
9            9              3.8   57190.0
10          10              4.0   63219.0
11          11              4.1   55795.0
12          12              4.1   56958.0
13          13              4.2   57082.0
14          14              4.6   61112.0
15          15              5.0   67939.0
16          16              5.2   66030.0
17          17              5.4   83089.0
18          18              6.0   81364.0
19          19              6.1   93941.0
20          20              6.9   91739.0
21          21              7.2   98274.0
22          22              8.0  101303.0
23          23              8.3  113813.0
24          24              8.8  109432.0
25          25              9.1  105583.0
26          26              9.6  116970.0
27          27              9.7  112636.0
28          28             10.4  122392.0
29          29             10.6  121873.0
```

```
df.head()
```

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

Next steps: Generate code with df | New interactive sheet

```
df.tail()
```

	Unnamed: 0	YearsExperience	Salary
25	25	9.1	105583.0
26	26	9.6	116970.0
27	27	9.7	112636.0
28	28	10.4	122392.0
29	29	10.6	121873.0

```
YearsExperience = df[['YearsExperience']]
Salary = df['Salary']

print("Input variable (YearsExperience):")
print(YearsExperience.head())
print("\nOutput variable (Salary):")
print(Salary.head())
```

```
Input variable (YearsExperience):
   YearsExperience
0              1.2
1              1.4
2              1.6
3              2.1
4              2.3

Output variable (Salary):
0    39344.0
1    46206.0
2    37732.0
3    43526.0
4    39892.0
Name: Salary, dtype: float64
```

Linear Regression using Scikit-learn

```
from sklearn.model_selection import tr
X = df[['YearsExperience']]   # Indepe
y = df['Salary']              # Depenc
# Split into 80% training and 20% test
X_train, X_test, y_train, y_test = tra
```

```
from sklearn.model_selection import tr
from sklearn.linear_model import Linea
```

```
# Create model
model = LinearRegression()
# Train model
model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

```
# Predict salaries for test set
y_pred = model.predict(X_test)
# Compare predictions with actual valu
print("Predicted salaries:", y_pred)
print("Actual salaries:", list(y_test)
```

```
Predicted salaries: [115791.21011287  71499.27809463 102597.86866063  75268.8
  55478.79204548  60190.69970699]
Actual salaries: [112636.0, 67939.0, 113813.0, 83089.0, 64446.0, 57190.0]
```

Performance Evaluation

```
from sklearn.metrics import mean_squar
# Calculate Mean Squared Error
mse = mean_squared_error(y_test, y_pre
print("Mean Squared Error:", mse)
# Calculate R-squared value
r2 = r2_score(y_test, y_pred)
print("R-squared:", r2)
```

```
Mean Squared Error: 49830096.855908394
R-squared: 0.9024461774180497
```

```
print("Slope (Coefficient):", model.coef_)
print("Intercept:", model.intercept_)
```

```
Slope (Coefficient): [9423.81532303]
Intercept: 24380.201479473704
```

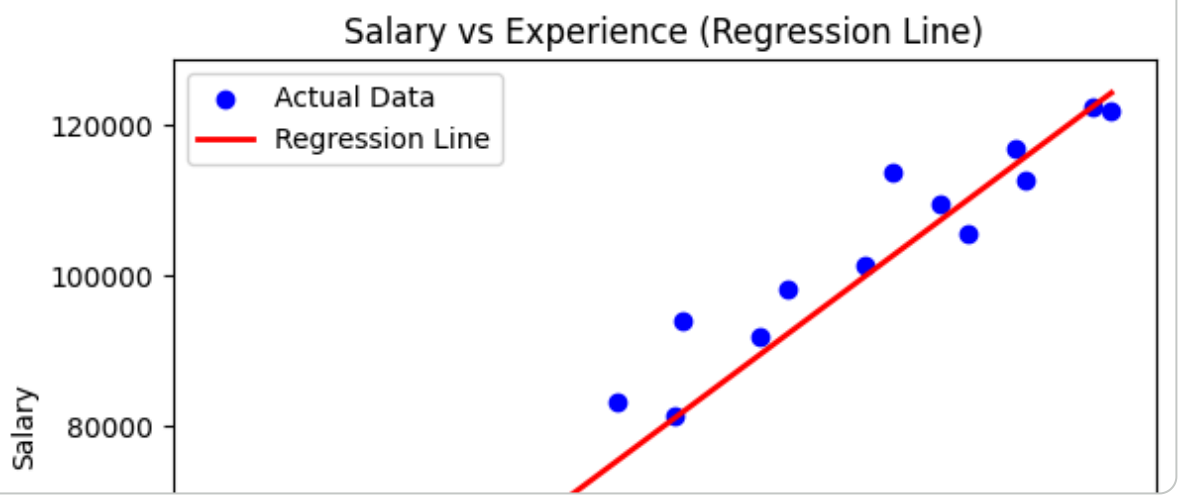
```
print("Predicted salaries:", y_pred)
print("Actual salaries:", list(y_test))
```

```
Predicted salaries: [115791.21011287  71499.27809463 102597.86866063  75268.8
  55478.79204548  60190.69970699]
Actual salaries: [112636.0, 67939.0, 113813.0, 83089.0, 64446.0, 57190.0]
```

```
from sklearn.metrics import mean_squared_error, r2_score
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print("MSE:", mse)
print("R²:", r2)
```

```
MSE: 49830096.855908394
R²: 0.9024461774180497
```

```
import matplotlib.pyplot as plt
# Scatter plot of actual data
plt.scatter(X, y, color='blue', label='Actual Data')
# Regression line
plt.plot(X, model.predict(X), color='red', linewidth=2, label='Regression Li
plt.xlabel("Years of Experience")
plt.ylabel("Salary")
plt.title("Salary vs Experience (Regression Line)")
plt.legend()
plt.show()
```




```
plt.scatter(y_test, y_pred, color='gre
plt.plot([y_test.min(), y_test.max()],
plt.xlabel("Actual Salary")
plt.ylabel("Predicted Salary")
plt.title("Actual vs Predicted Salarie
plt.show()
```