```
# TASK 1
# 1. Import thư viện cần thiết
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# 2. Định nghĩa tên file
filename = "50_Startups.csv"

# 3. Kiểm tra nếu file chưa tồn tại thì yêu cầu tải lên
if not os.path.exists(filename):
    from google.colab import files
    uploaded = files.upload()
    filename = list(uploaded.keys())[0]  # Lấy tên file vừa tải lên

# 4. Đọc file CSV vào DataFrame
df = pd.read_csv(filename)

# 5. Loại bỏ các cột chứa dữ liệu không phải số
df_numeric = df.select_dtypes(include=[np.number])  # Chỉ giữ lại các cột số

# 6. Hiển thị 5 dòng đầu tiên để kiểm tra dữ liệu
print("\n📌 Dữ liệu số ban đầu:")
print(df_numeric.head())

# 7. Tính toán các thống kê cơ bản
summary = df_numeric.describe()
print("\n📌 Thống kê dữ liệu:\n", summary)

# 8. Vẽ scatter plot giữa từng feature và biến phụ thuộc y
y = df_numeric.iloc[:, -1]  # Cột cuối cùng là biến phụ thuộc (y)
features = df_numeric.columns[:-1]  # Các cột còn lại là feature

for feature in features:
    plt.scatter(df_numeric[feature], y)
    plt.xlabel(feature)
    plt.ylabel("Output (y)")
    plt.title(f'Scatter plot: {feature} vs Output')
    plt.show()

# 9. Vẽ histogram của từng feature
for feature in df_numeric.columns:
    mean = df_numeric[feature].mean()
    var = df_numeric[feature].var()
    plt.hist(df_numeric[feature], bins=20, alpha=0.7, color='blue')
    plt.title(f'Histogram of {feature} (mean={mean:.1f}, var={var:.1f})')
    plt.xlabel(feature)
    plt.ylabel("Frequency")
    plt.show()

# 10. Tính hệ số tương quan giữa các biến
correlations = df_numeric.corr()
print("\n📌 Hệ số tương quan giữa y và các feature:")
print(correlations.iloc[-1, :-1])  # In hệ số tương quan của y với các feature

# 11. Chuẩn hóa dữ liệu
# a. Minmax scaling (đưa dữ liệu về khoảng [0,1])
X_minmax = (df_numeric.iloc[:, :-1] - df_numeric.iloc[:, :-1].min()) / (df_numeric.iloc[:, :-1].max() - df_numeric.iloc[:, :-1].min())

# b. Standardization (chuẩn hóa sao cho mean=0, variance=1)
X_standardized = (df_numeric.iloc[:, :-1] - df_numeric.iloc[:, :-1].mean()) / df_numeric.iloc[:, :-1].std()

# 12. Chuyển đổi dữ liệu thành X1 (min=1, max=10)
X1 = df_numeric.iloc[:, :-1].apply(lambda x: 1 + 9 * (x - x.min()) / (x.max() - x.min()))

# 13. Chuyển đổi dữ liệu thành X2 bằng chuẩn hóa (standardize)
X2 = X_standardized.copy()

# 14. In kết quả để kiểm tra
print("\n📌 Dữ liệu sau MinMax Scaling:\n", X_minmax.head())
print("\n📌 Dữ liệu sau Standardization:\n", X_standardized.head())
print("\n📌 Dữ liệu sau chuyển đổi thành X1:\n", X1.head())
print("\n📌 Dữ liệu sau chuẩn hóa thành X2:\n", X2.head())

```

Choose Files 50_Startups.csv

- **50_Startups.csv**(text/csv) - 2436 bytes, last modified: 2/14/2025 - 100% done

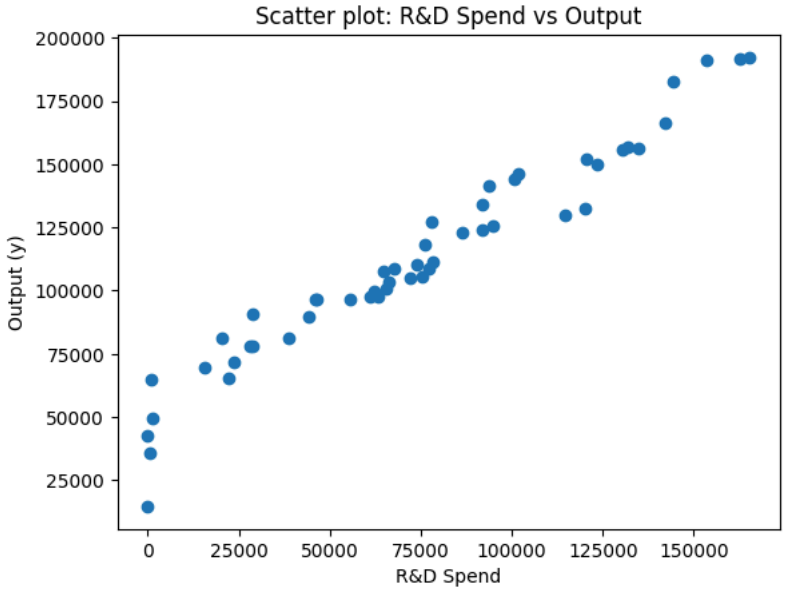
```
Saving 50_Startups.csv to 50_Startups.csv
```

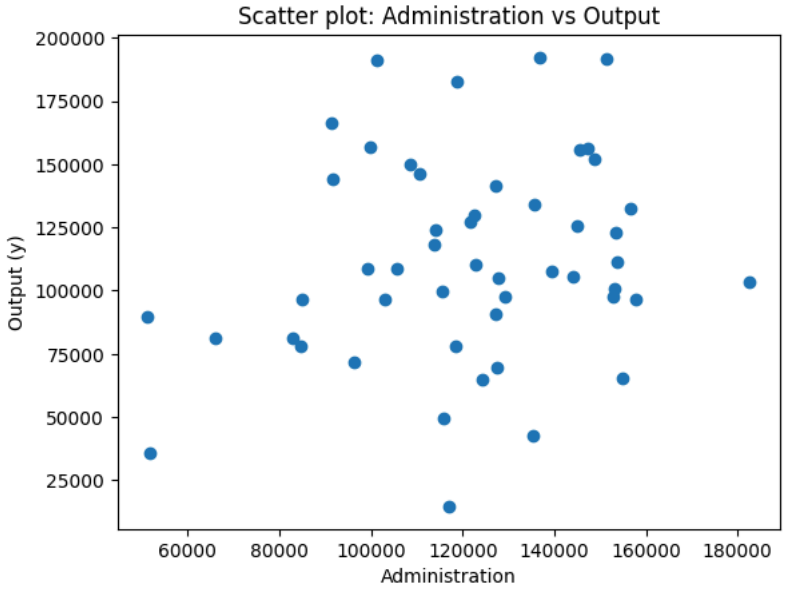
📌 Dữ liệu số ban đầu:

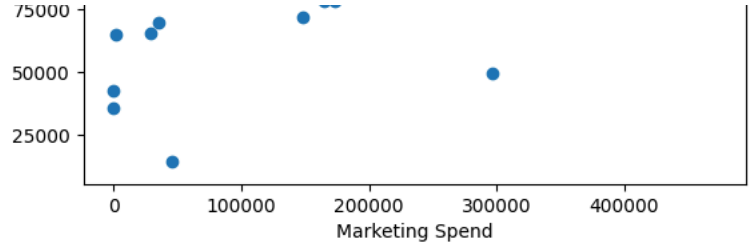
	R&D Spend	Administration	Marketing Spend	Profit
0	165349.20	136897.80	471784.10	192261.83
1	162597.70	151377.59	443898.53	191792.06
2	153441.51	101145.55	407934.54	191050.39
3	144372.41	118671.85	383199.62	182901.99
4	142107.34	91391.77	366168.42	166187.94

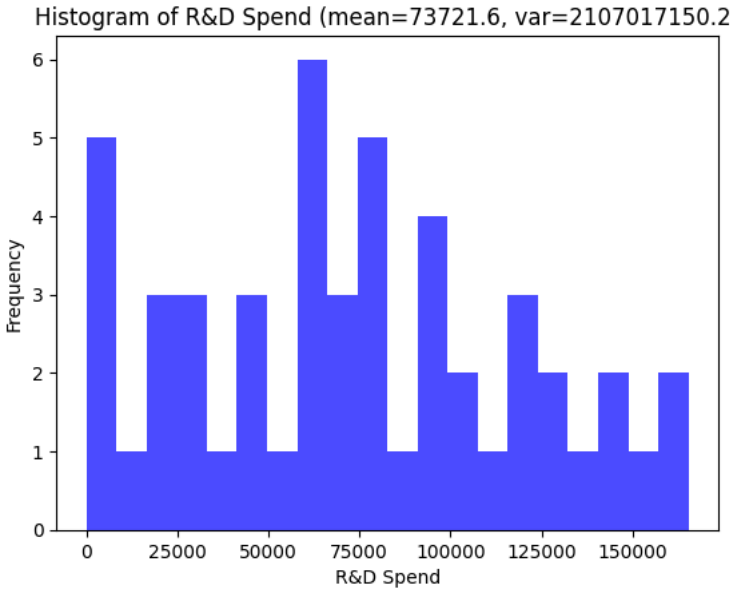
📌 Thống kê dữ liệu:

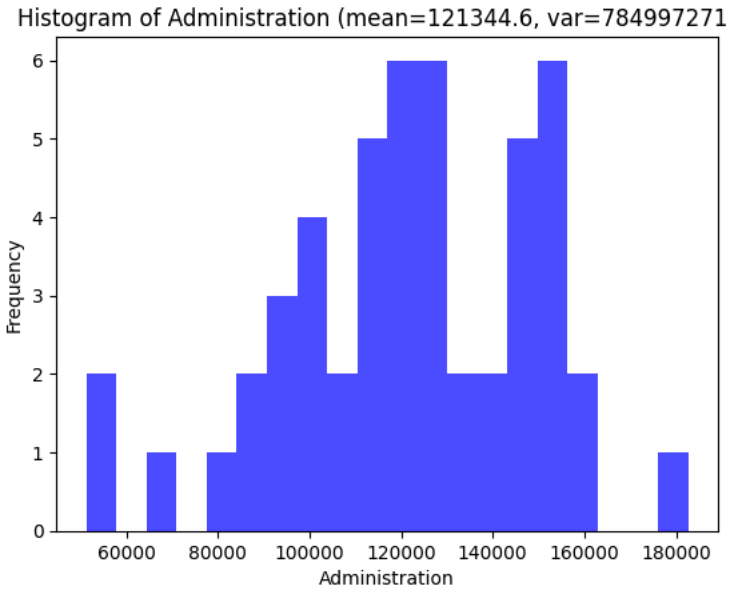
	R&D Spend	Administration	Marketing Spend	Profit
count	50.000000	50.000000	50.000000	50.000000
mean	73721.615600	121344.639600	211025.097800	112012.639200
std	45902.256482	28017.802755	122290.310726	40306.180338
min	0.000000	51283.140000	0.000000	14681.400000
25%	39936.370000	103730.875000	129300.132500	90138.902500
50%	73051.080000	122699.795000	212716.240000	107978.190000
75%	101602.800000	144842.180000	299469.085000	139765.977500
max	165349.200000	182645.560000	471784.100000	192261.830000

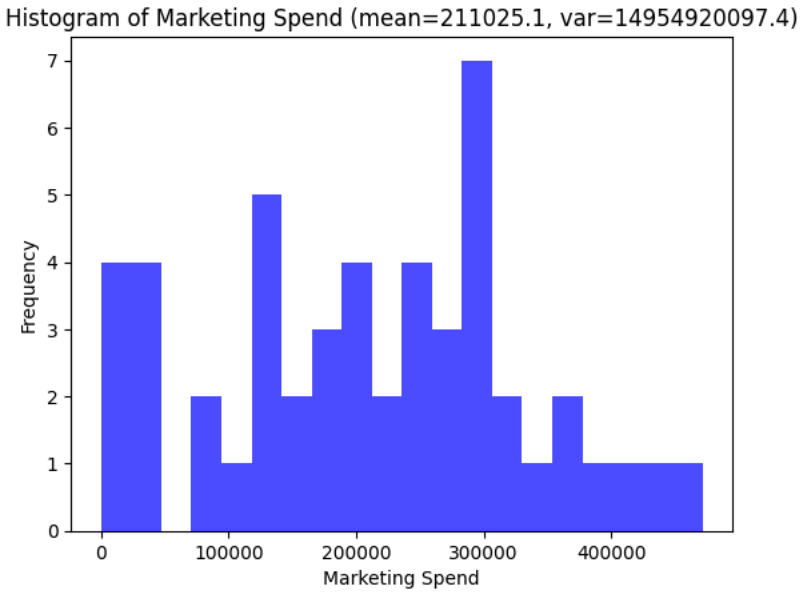




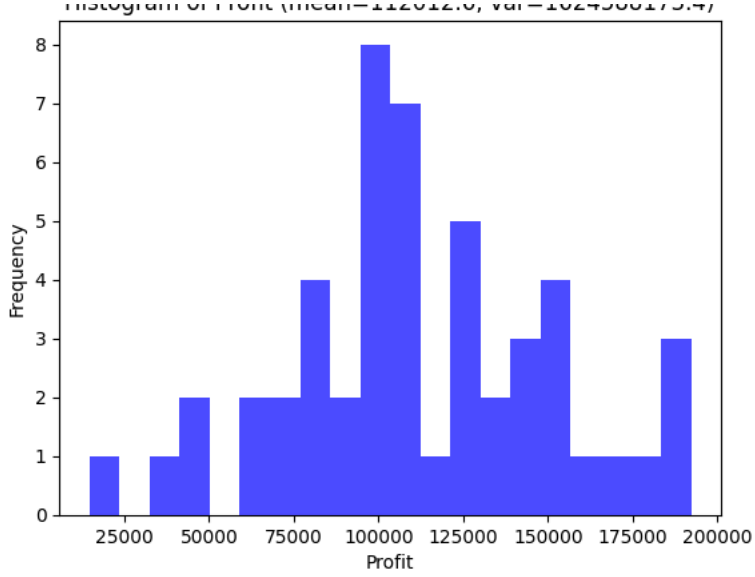








Histogram of Profit (mean=112012.6, var=1624588173.4)



📌 Hệ số tương quan giữa y và các feature:
```
R&D Spend          0.972900
Administration     0.200717
Marketing Spend    0.747766
Name: Profit, dtype: float64
```

📌 Dữ liệu sau MinMax Scaling:

	R&D Spend	Administration	Marketing Spend
0	1.000000	0.651744	1.000000
1	0.983359	0.761972	0.940893
2	0.927985	0.379579	0.864664
3	0.873136	0.512998	0.812235
4	0.859438	0.305328	0.776136

📌 Dữ liệu sau Standardization:

	R&D Spend	Administration	Marketing Spend
0	1.996146	0.555117	2.132295
1	1.936203	1.071924	1.904267
2	1.736731	-0.720938	1.610180
3	1.539157	-0.095396	1.407916
4	1.489812	-1.069066	1.268648

📌 Dữ liệu sau chuyển đổi thành X1:

	R&D Spend	Administration	Marketing Spend
0	10.000000	6.865695	10.000000
1	9.850235	7.857746	9.468040
2	9.351861	4.416211	8.781972
3	8.858228	5.616986	8.310116
4	8.734940	3.747952	7.985220

📌 Dữ liệu sau chuẩn hóa thành X2:

	R&D Spend	Administration	Marketing Spend
0	1.996146	0.555117	2.132295
1	1.936203	1.071924	1.904267
2	1.736731	-0.720938	1.610180
3	1.539157	-0.095396	1.407916
4	1.489812	-1.069066	1.268648