CHRIST
(DEEMED TO BE UNIVERSITY)
DELHI-NCR, INDIA

AI/ML Programming

MCA-475

Assignment – 04

BY

HIMANSHU HEDA (24225013)

SUBMITTED TO

Dr. Manjula Shannhog

SCHOOL OF SCIENCES

2025-26

Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import date
```

```
df = pd.read_csv('./Dataset/50_Startups.csv')
```

```
df.head()
```

✓ 0.0s Python

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

```
df.tail()
```

✓ 0.0s Python

	R&D Spend	Administration	Marketing Spend	State	Profit
45	1000.23	124153.04	1903.93	New York	64926.08
46	1315.46	115816.21	297114.46	Florida	49490.75
47	0.00	135426.92	0.00	California	42559.73
48	542.05	51743.15	0.00	New York	35673.41
49	0.00	116983.80	45173.06	California	14681.40

```
df.info()
```

✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   R&D Spend        50 non-null     float64
 1   Administration   50 non-null     float64
 2   Marketing Spend  50 non-null     float64
 3   State            50 non-null     object 
 4   Profit           50 non-null     float64
dtypes: float64(4), object(1)
memory usage: 2.1+ KB
```

```
df.dtypes
```
✓ 0.0s Python

```
R&D Spend          float64
Administration     float64
Marketing Spend    float64
State               object
Profit             float64
dtype: object
```

Generate + Code + Markdown

```
df.rename(columns={'Administration': 'Administration (Admin)', 'Marketing Spend': 'Advertisement Expenses'}, inplace=True)
df.head()
```
✓ 0.0s Python

	R&D Spend	Administration (Admin)	Advertisement Expenses	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

```
df.sample(5)
```
✓ 0.0s Python

	R&D Spend	Administration (Admin)	Advertisement Expenses	State	Profit
2	153441.51	101145.55	407934.54	Florida	191050.39
25	64664.71	139553.16	137962.62	California	107404.34
18	91749.16	114175.79	294919.57	Florida	124266.90
17	94657.16	145077.58	282574.31	New York	125370.37
11	100671.96	91790.61	249744.55	California	144259.40

```
df.shape
```
✓ 0.0s Python

```
(50, 5)
```

```
df.columns
```
✓ 0.0s Python

```
Index(['R&D Spend', 'Administration (Admin)', 'Advertisement Expenses',
       'State', 'Profit'],
      dtype='object')
```

```
df.isnull().sum()
```
✓ 0.0s Python

```
R&D Spend                0
Administration (Admin)   0
Advertisement Expenses   0
State                    0
Profit                   0
dtype: int64
```

```
df.notnull().sum()
```
✓ 0.0s Python

```
R&D Spend                50
Administration (Admin)   50
Advertisement Expenses   50
State                    50
Profit                   50
dtype: int64
```

```
df.nunique
```
✓ 0.0s Python

```
df.nunique
```

[20] ✓ 0.0s Python

```
<bound method DataFrame.nunique of     R&D Spend  Administration (Admin)  Advertisement Expenses       State  \
0   165349.20               136897.80               471784.10    New York
1   162597.70               151377.59               443898.53  California
2   153441.51               101145.55               407934.54     Florida
3   144372.41               118671.85               383199.62    New York
4   142107.34                91391.77               366168.42     Florida
5   131876.90                99814.71               362861.36    New York
6   134615.46               147198.87               127716.82  California
7   130298.13               145530.06               323876.68     Florida
8   120542.52               148718.95               311613.29    New York
9   123334.88               108679.17               304981.62  California
10  101913.08               110594.11               229160.95     Florida
11  100671.96                91790.61               249744.55  California
12   93863.75               127320.38               249839.44     Florida
13   91992.39               135495.07               252664.93  California
14  119943.24               156547.42               256512.92     Florida
15  114523.61               122616.84               261776.23    New York
16   78013.11               121597.55               264346.06  California
17   94657.16               145077.58               282574.31    New York
18   91749.16               114175.79               294919.57     Florida
19   86419.70               153514.11                    0.00    New York
20   76253.86               113867.30               298664.47  California
```

```
df.sort_values('Profit', ascending=False).head()
```

[21] ✓ 0.0s Python

	R&D Spend	Administration (Admin)	Advertisement Expenses	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

```
df.describe(include='all')
```

[22] ✓ 0.0s Python

	R&D Spend	Administration (Admin)	Advertisement Expenses	State	Profit
count	50.000000	50.000000	50.000000	50	50.000000
unique	NaN	NaN	NaN	3	NaN
top	NaN	NaN	NaN	New York	NaN
freq	NaN	NaN	NaN	17	NaN

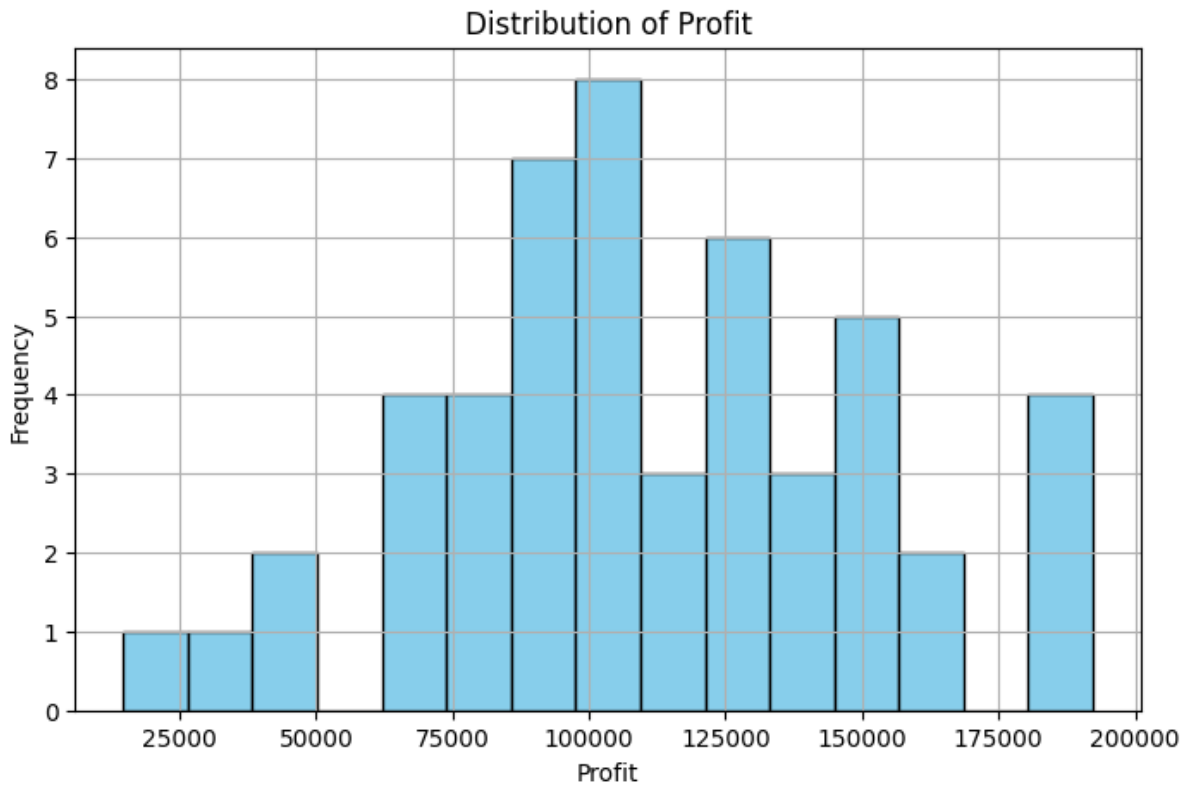
```
df.describe(include='all')
```

✓ 0.0s Python

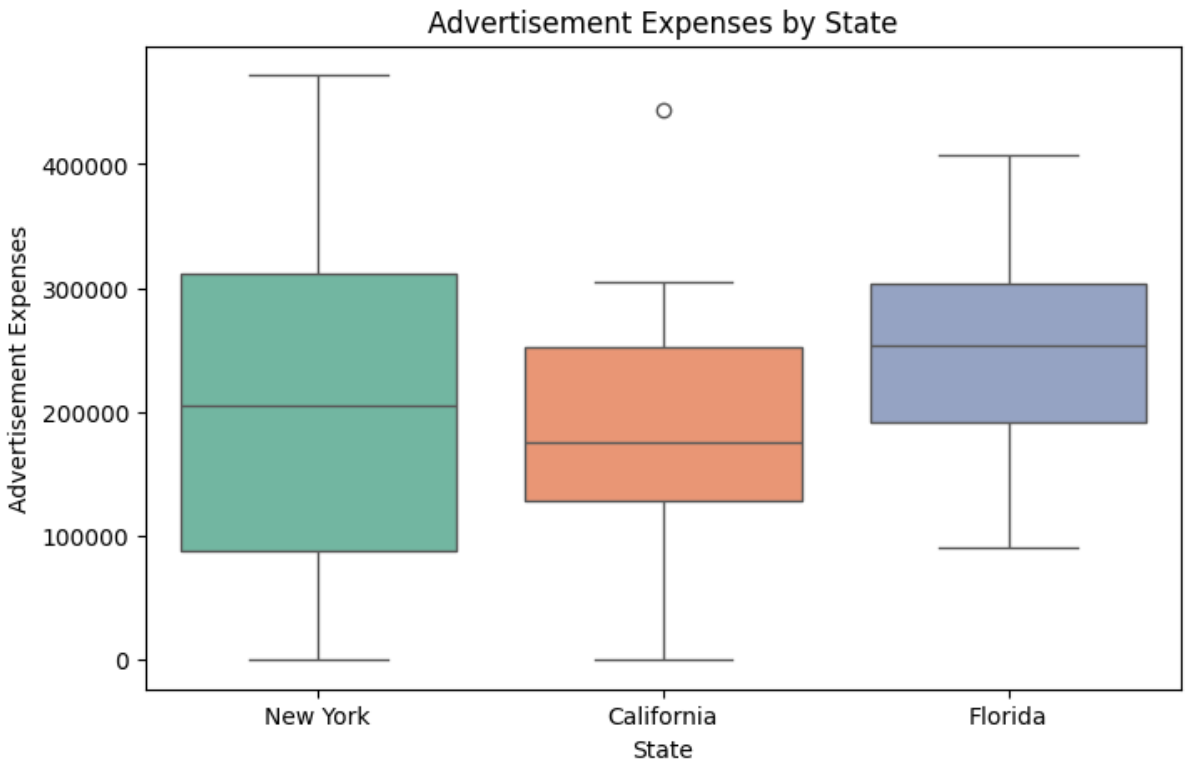
	R&D Spend	Administration (Admin)	Advertisement Expenses	State	Profit
count	50.000000	50.000000	50.000000	50	50.000000
unique	NaN	NaN	NaN	3	NaN
top	NaN	NaN	NaN	New York	NaN
freq	NaN	NaN	NaN	17	NaN
mean	73721.615600	121344.639600	211025.097800	NaN	112012.639200
std	45902.256482	28017.802755	122290.310726	NaN	40306.180338
min	0.000000	51283.140000	0.000000	NaN	14681.400000
25%	39936.370000	103730.875000	129300.132500	NaN	90138.902500
50%	73051.080000	122699.795000	212716.240000	NaN	107978.190000
75%	101602.800000	144842.180000	299469.085000	NaN	139765.977500
max	165349.200000	182645.560000	471784.100000	NaN	192261.830000

Graphical Presentation :

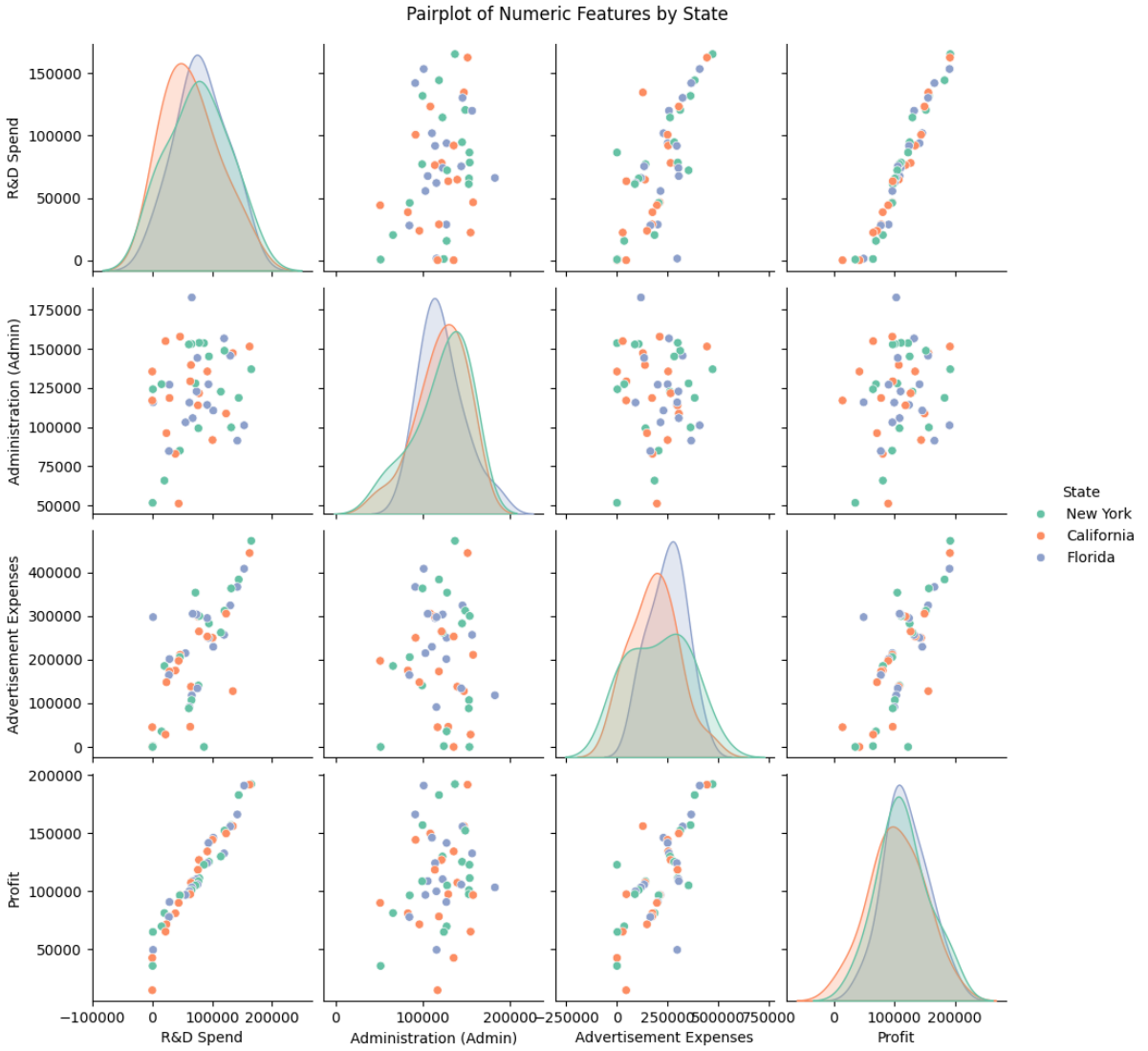
```
# 1. Histogram of Profit
plt.figure(figsize=(8,5))
df['Profit'].hist(bins=15, color='skyblue', edgecolor='black')
plt.title('Distribution of Profit')
plt.xlabel('Profit')
plt.ylabel('Frequency')
plt.show()
```



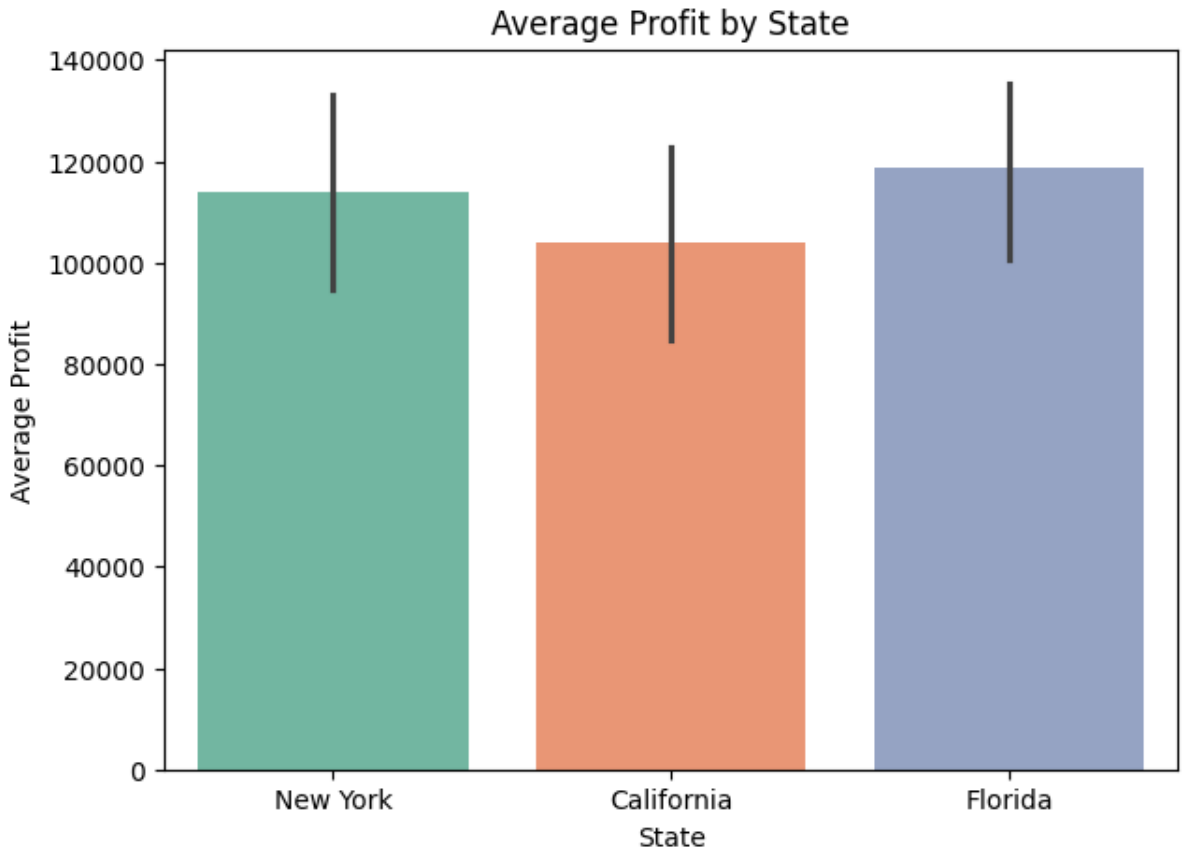

```
# 2. Boxplot of Advertisement Expenses by State
plt.figure(figsize=(8,5))
sns.boxplot(data=df, x='State', y='Advertisement Expenses', palette='Set2')
plt.title('Advertisement Expenses by State')
plt.xlabel('State')
plt.ylabel('Advertisement Expenses')
plt.show()
```



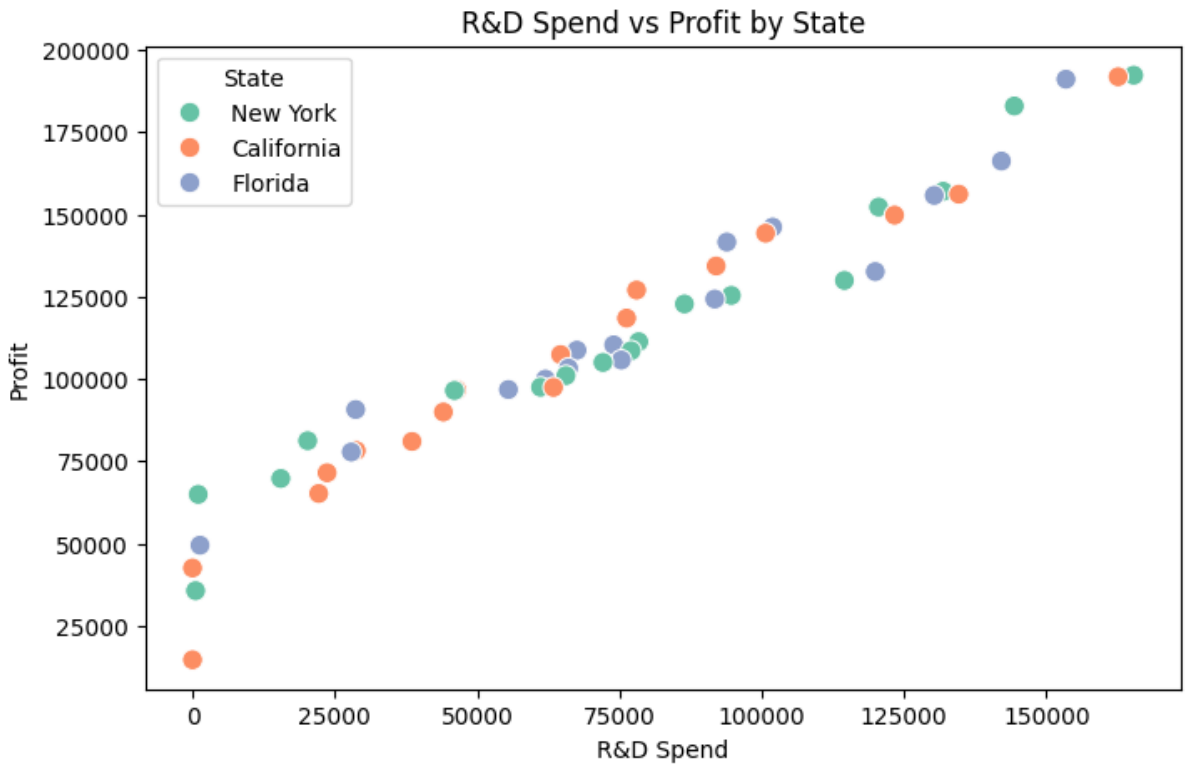

```
# 3. Pairplot of numeric features
sns.pairplot(df, vars=['R&D Spend', 'Administration (Admin)', 'Advertisement
Expenses', 'Profit'], hue='State', palette='Set2')
plt.suptitle('Pairplot of Numeric Features by State', y=1.02)
plt.show()
```



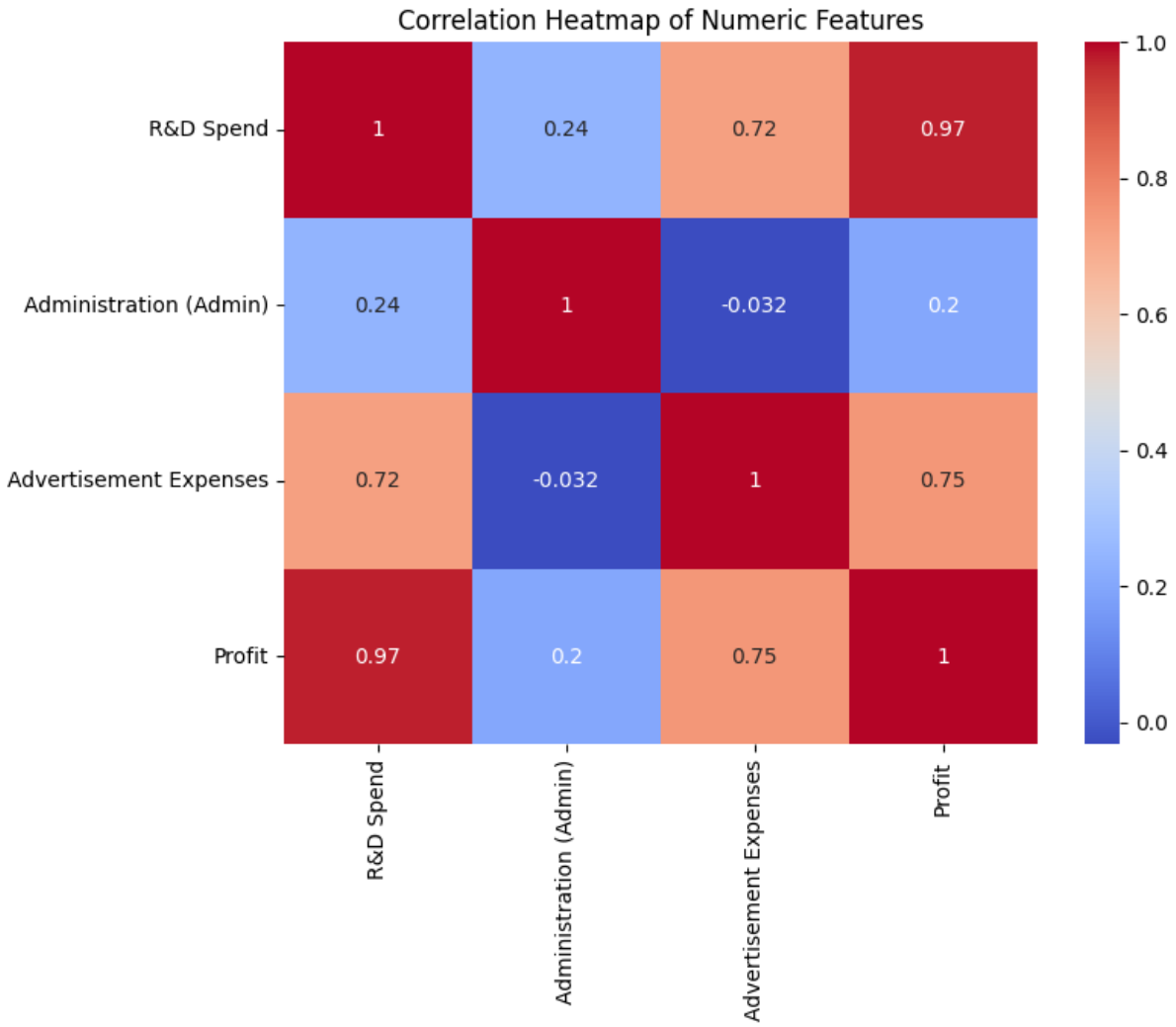

```
# 4. Barplot of average Profit by State
plt.figure(figsize=(7,5))
sns.barplot(data=df, x='State', y='Profit', estimator=np.mean, palette='Set2')
plt.title('Average Profit by State')
plt.xlabel('State')
plt.ylabel('Average Profit')
plt.show()
```



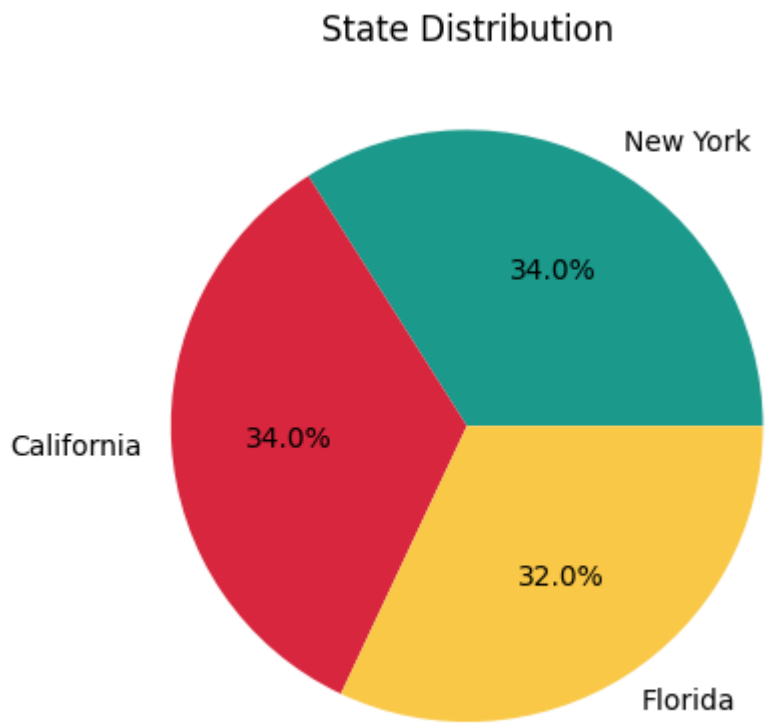

```
# 5. Scatterplot of R&D Spend vs Profit
plt.figure(figsize=(8,5))
sns.scatterplot(data=df, x='R&D Spend', y='Profit', hue='State',
palette='Set2', s=80)
plt.title('R&D Spend vs Profit by State')
plt.xlabel('R&D Spend')
plt.ylabel('Profit')
plt.legend(title='State')
plt.show()
```



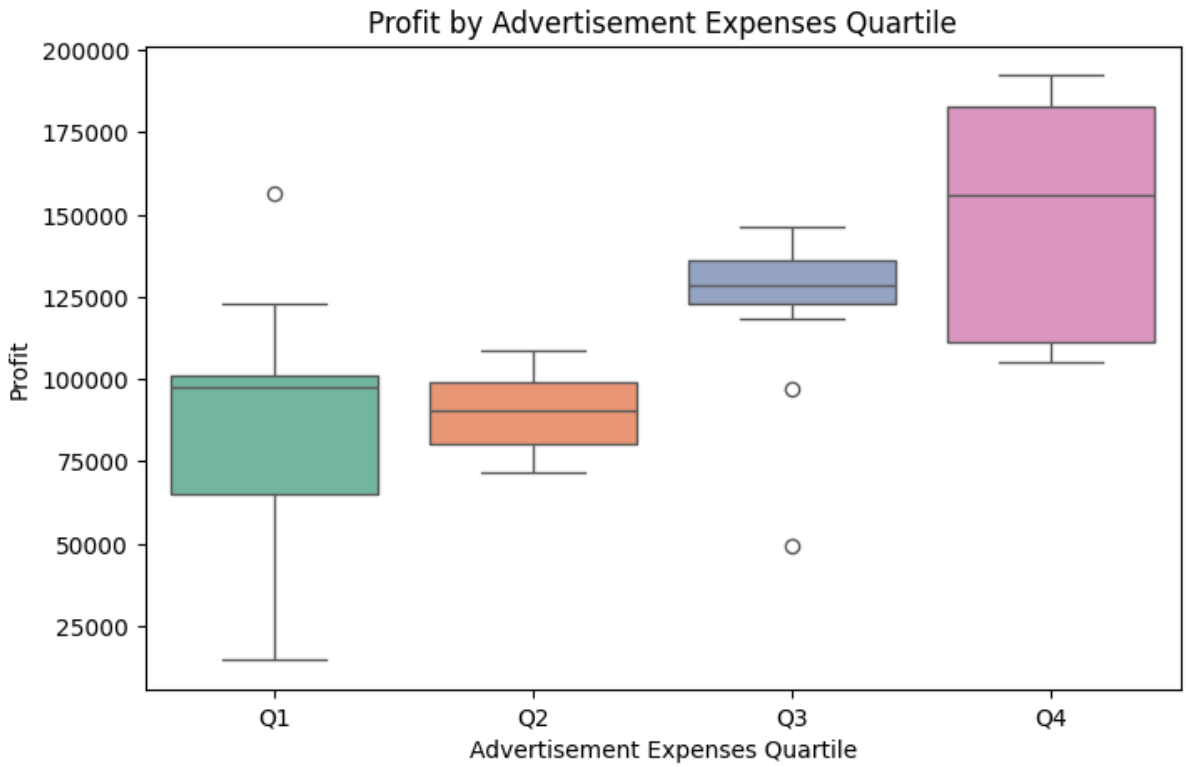

```
# 6. Correlation heatmap
plt.figure(figsize=(8,6))
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap of Numeric Features')
plt.show()
```



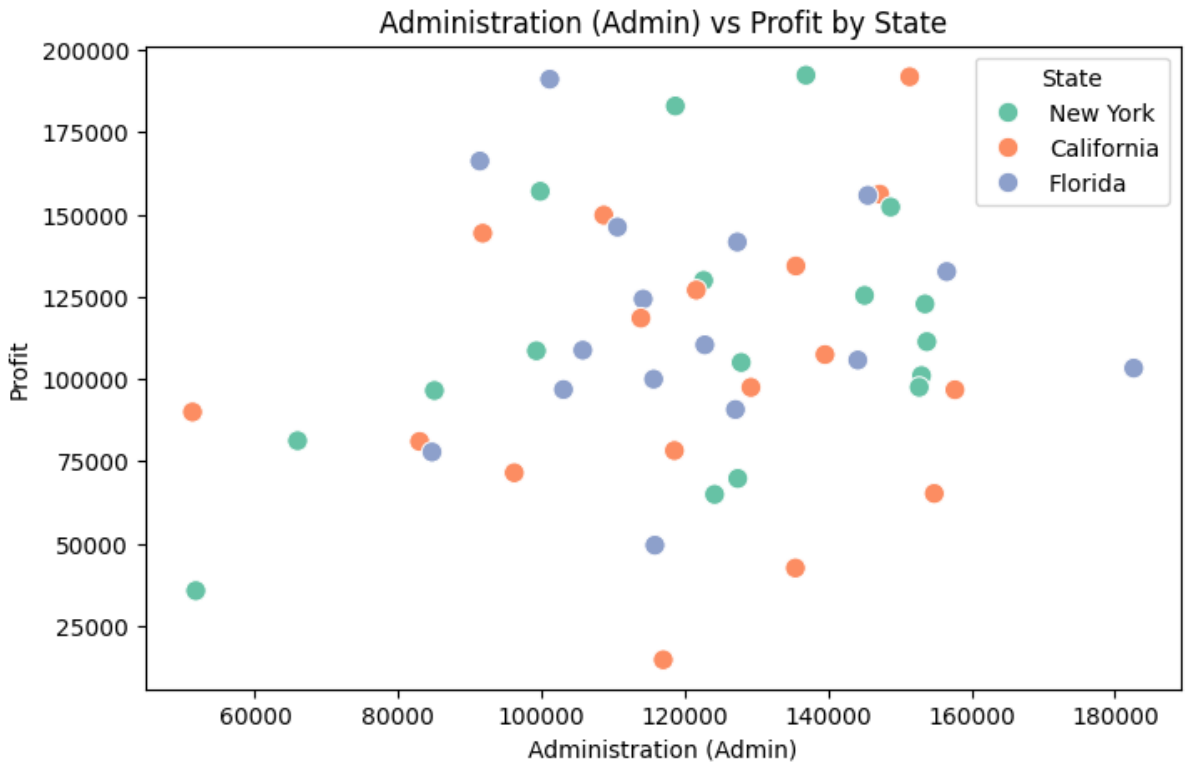

```
# 7. Pie chart of State distribution
df['State'].value_counts().plot(kind='pie', autopct='%1.1f%%',
colors=['#1b998b','#d7263d','#f9c846'])
plt.title('State Distribution')
plt.ylabel('')
plt.show()
```




```
# 8. Boxplot of Profit by Advertisement Expenses quartiles
df['Ad_Quartile'] = pd.qcut(df['Advertisement Expenses'], 4,
labels=['Q1','Q2','Q3','Q4'])
plt.figure(figsize=(8,5))
sns.boxplot(data=df, x='Ad_Quartile', y='Profit', palette='Set2')
plt.title('Profit by Advertisement Expenses Quartile')
plt.xlabel('Advertisement Expenses Quartile')
plt.ylabel('Profit')
plt.show()
df.drop('Ad_Quartile', axis=1, inplace=True)
```




```
# 9. Scatterplot of Administration (Admin) vs Profit
plt.figure(figsize=(8,5))
sns.scatterplot(data=df, x='Administration (Admin)', y='Profit', hue='State',
palette='Set2', s=80)
plt.title('Administration (Admin) vs Profit by State')
plt.xlabel('Administration (Admin)')
plt.ylabel('Profit')
plt.legend(title='State')
plt.show()
```




```
# 10. Top 10 startups by Advertisement Expenses
print('Top 10 Startups by Advertisement Expenses:')
print(df.sort_values('Advertisement Expenses', ascending=False).head(10))
```

✓ 0.0s Python

```
Top 10 Startups by Advertisement Expenses:
    R&D Spend  Administration (Admin)  Advertisement Expenses       State  \
0   165349.20               136897.80               471784.10    New York
1   162597.70               151377.59               443898.53  California
2   153441.51               101145.55               407934.54     Florida
3   144372.41               118671.85               383199.62    New York
4   142107.34                91391.77               366168.42     Florida
5   131876.90                99814.71               362861.36    New York
27   72107.60               127864.55               353183.81    New York
7   130298.13               145530.06               323876.68     Florida
8   120542.52               148718.95               311613.29    New York
9   123334.88               108679.17               304981.62  California

       Profit
0   192261.83
1   191792.06
2   191050.39
3   182901.99
4   166187.94
5   156991.12
```

```
       Profit
0   192261.83
1   191792.06
2   191050.39
3   182901.99
4   166187.94
5   156991.12
27  105008.31
7   155752.60
8   152211.77
9   149759.96
```