Problem statement ¶

Prepare a prediction model for profit of 50_startups data.

Importing the libraries

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sn
from statsmodels.graphics.regressionplot
s import influence_plot
import statsmodels.formula.api as smf
import numpy as np
from sklearn.preprocessing import Standa
rdScaler, MinMaxScaler
```

We can just peek into few data points by using head function of pandas. By default, head function return top 5 values

Data insights

In [4]:

```
startups.shape
```

Out[4]:

```
(50, 5)
```

In [5]:

```
startups.columns
```

Out[5]:

```
Index(['R&D Spend', 'Administration', 'M
arketing Spend', 'State', 'Profit'], dty
pe='object')
```

Loading dataset

In [2]:

```
startups = pd.read_csv("/kaggle/input/startup-logistic-regression/50_Startups.csv")
```

In [3]:

```
startups.head()
```

Out[3]:

	R&D Spend	Administration	Marketing Spend	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

```
---
 0   R&D Spend         50 non-null     fl
oat64
 1   Administration    50 non-null     fl
oat64
 2   Marketing Spend   50 non-null     fl
oat64
 3   State             50 non-null     ob
ject
 4   Profit            50 non-null     fl
oat64
dtypes: float64(4), object(1)
memory usage: 2.1+ KB
```

Observations :-

1. We can see that R&D spend, Administration, Marketing Spend and Profit consists of floating point data type values and State has object type values.
2. We can also see that all 21 observations are non null and hence we don't have any missing values

Observations :-

1. The dataset contains data about 50 startups. It has 5 columns: "R&D Spend", "Administration", "Marketing Spend", "State", "Profit".
2. The first 3 columns indicate how much each startup spends on Research and Development, how much they spend on Marketing, and how much they spend on Administration cost.
3. The state column indicates which state the startup is based in and the last column states the profit made by the startup.

In [6]:

```
startups.info()
```

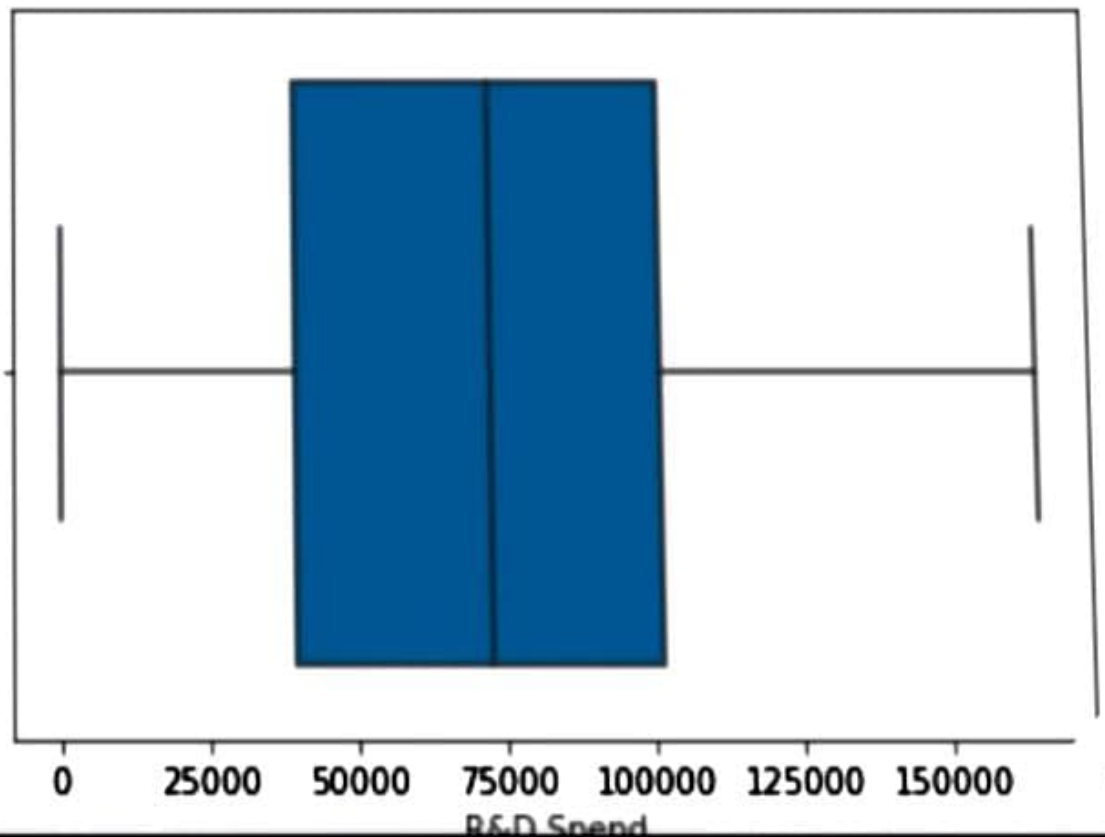
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dt
ype
---  ------           --------------  -
```

```
sn.boxplot(startups['R&D Spend'])
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

Out[11]:

```
<AxesSubplot:xlabel='R&D Spend'>
```



In [9]:

```
startups['Profit'].unique()
```

Out[9]:

```
array([192261.83, 191792.06, 191050.39,
182901.99, 166187.94, 156991.12,
       156122.51, 155752.6 , 152211.77,
149759.96, 146121.95, 144259.4 ,
       141585.52, 134307.35, 132602.65,
129917.04, 126992.93, 125370.37,
       124266.9 , 122776.86, 118474.03,
111313.02, 110352.25, 108733.99,
       108552.04, 107404.34, 105733.54,
105008.31, 103282.38, 101004.64,
        99937.59,  97483.56,  97427.84,
96778.92,  96712.8 ,  96479.51,
        90708.19,  89949.14,  81229.06,
81005.76,  78239.91,  77798.83,
        71498.49,  69758.98,  65200.33,
64926.08,  49490.75,  42559.73,
        35673.41,  14681.4 ])
```

In [7]:

```
startups[startups.duplicated()]
```

Out[7]:

	R&D Spend	Administration	Marketing Spend	State	Profit

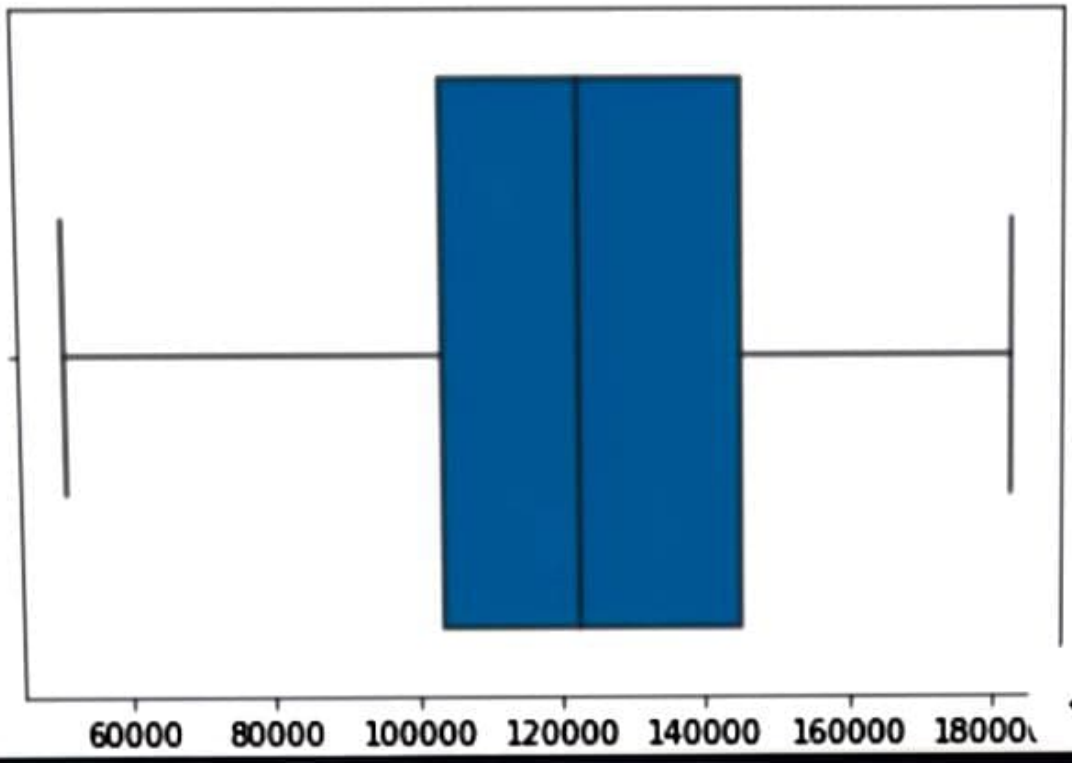
We don't have any duplicate values in our dataset. If duplicates values would have been present we would have to delete it.

```
sn.boxplot(startups['Administration'])
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

Out[12]:

```
<AxesSubplot:xlabel='Administration'>
```

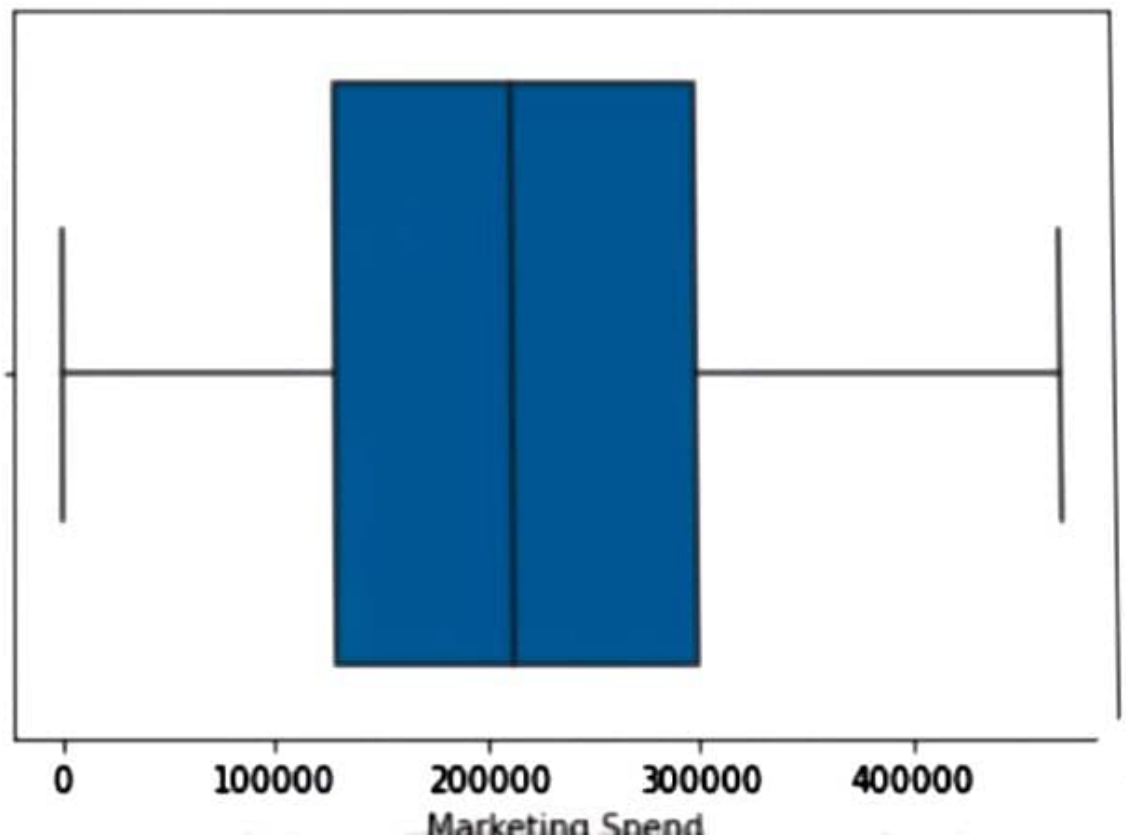


```
sn.boxplot(startups['Marketing Spend'])
```

```
/opt/conda/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

Out[13]:

```
<AxesSubplot:xlabel='Marketing Spend'>
```

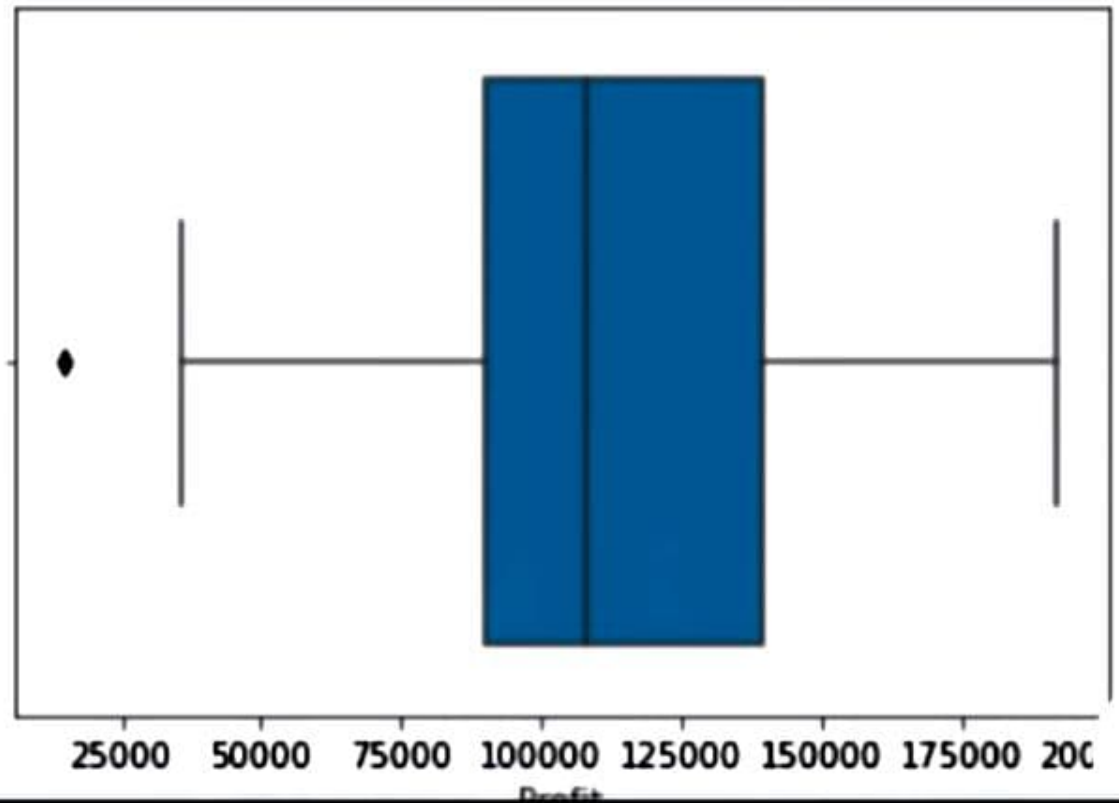


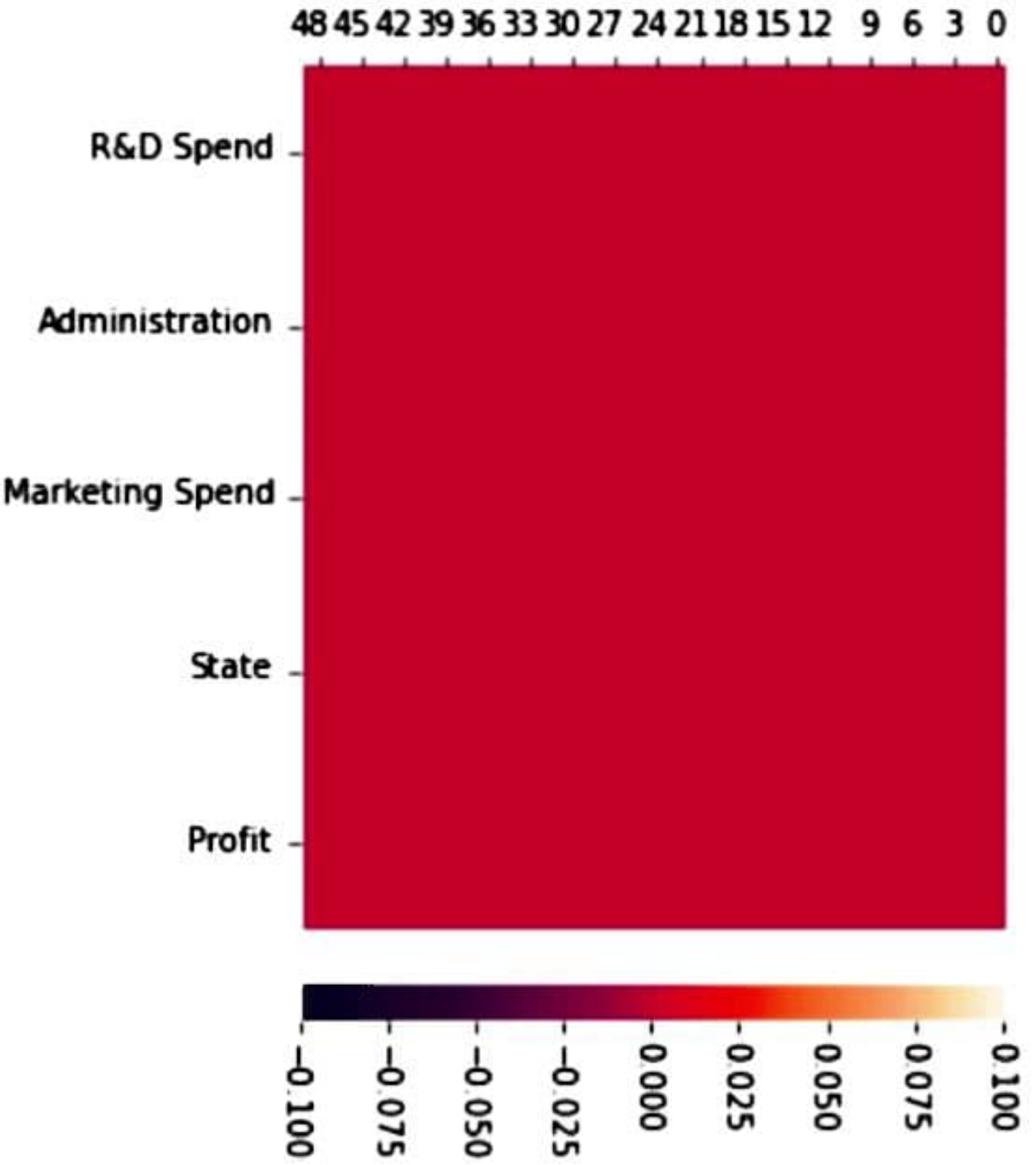
```
sn.boxplot(startups['Profit'])
```

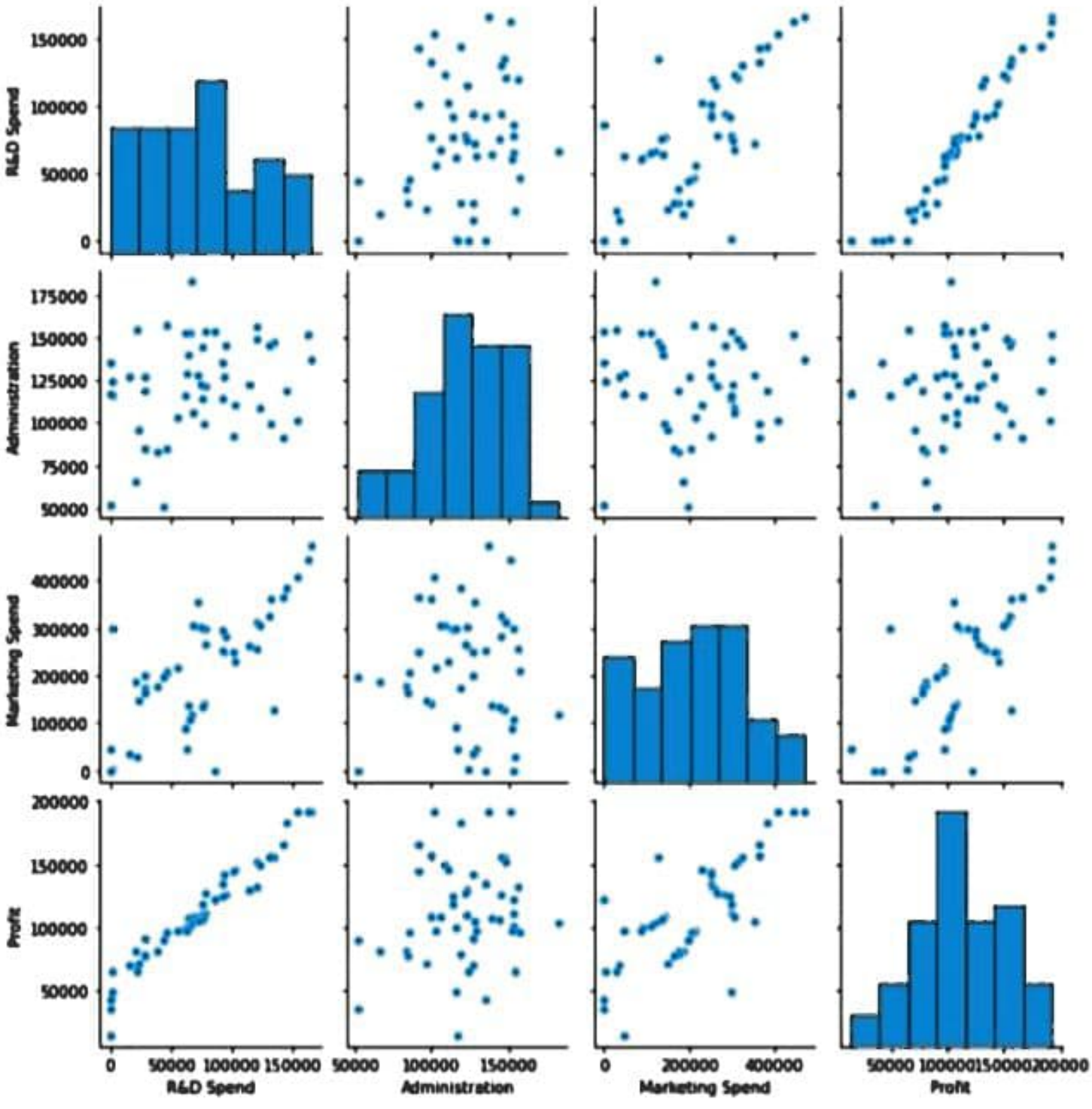
```
/opt/conda/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

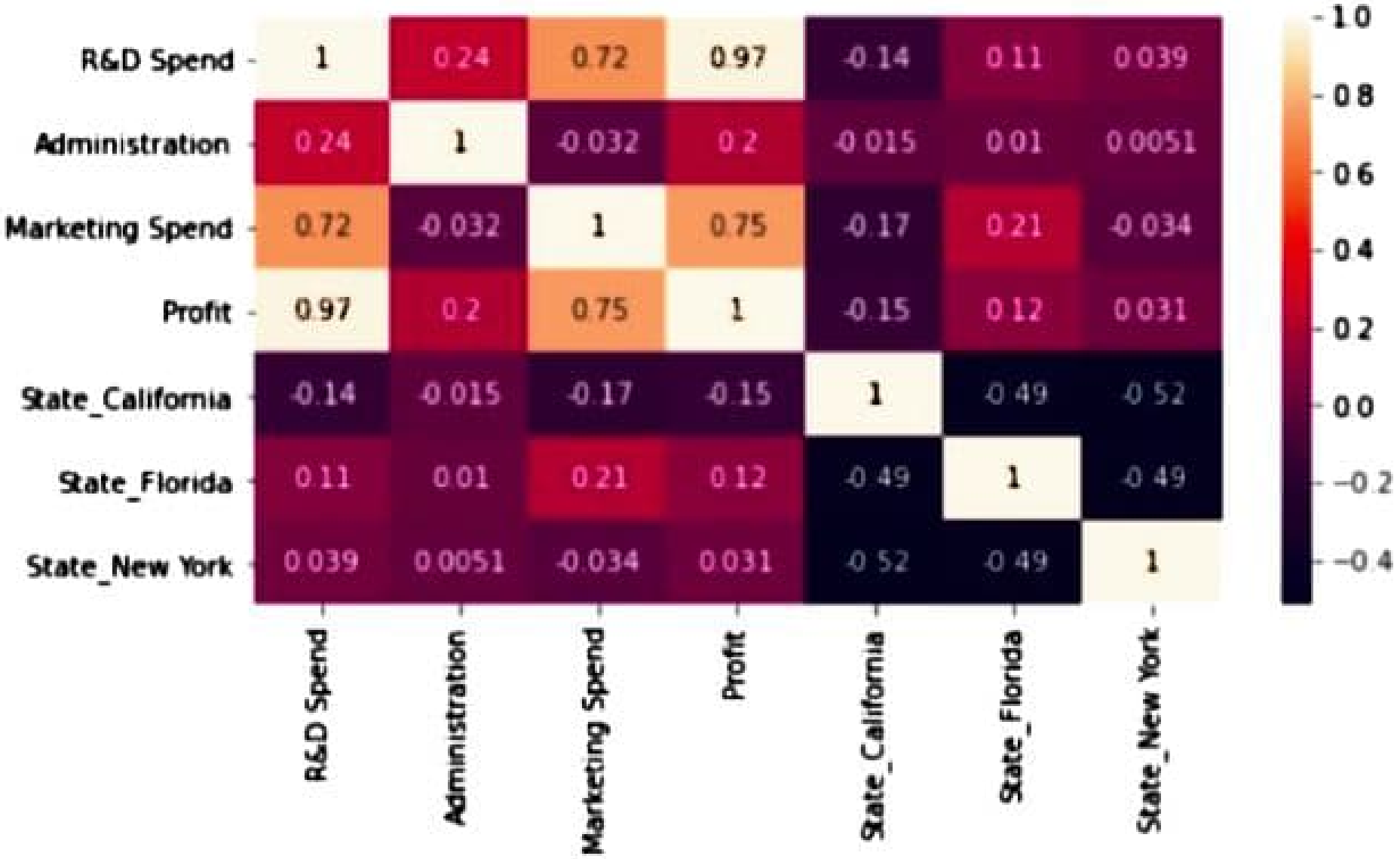
Out[14]:

```
<AxesSubplot:xlabel='Profit'>
```




48 45 42 39 36 33 30 27 24 21 18 15 12 9 6 3 0
R&D Spend
Administration
Marketing Spend
State
Profit
−0.100 −0.075 −0.050 −0.025 0.000 0.025 0.050 0.075 0.100


R&D Spend
Administration
Marketing Spend
Profit
150000
100000
50000
0
175000
150000
125000
100000
75000
50000
400000
300000
200000
100000
0
200000
150000
100000
50000
0
50000
100000
150000
0
200000
400000
50000
100000
150000
200000


R&D Spend
Administration
Marketing Spend
Profit
State_California
State_Florida
State_New York
1 0.24 0.72 0.97 -0.14 0.11 0.039
0.24 1 -0.032 0.2 -0.015 0.01 0.0051
0.72 -0.032 1 0.75 -0.17 0.21 -0.034
0.97 0.2 0.75 1 -0.15 0.12 0.031
-0.14 -0.015 -0.17 -0.15 1 -0.49 -0.52
0.11 0.01 0.21 0.12 -0.49 1 -0.49
0.039 0.0051 -0.034 0.031 -0.52 -0.49 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Sales Dashboard

Introduction | Overview | Market Analysis | Product Shipments

KEY METRICS SUMMARY

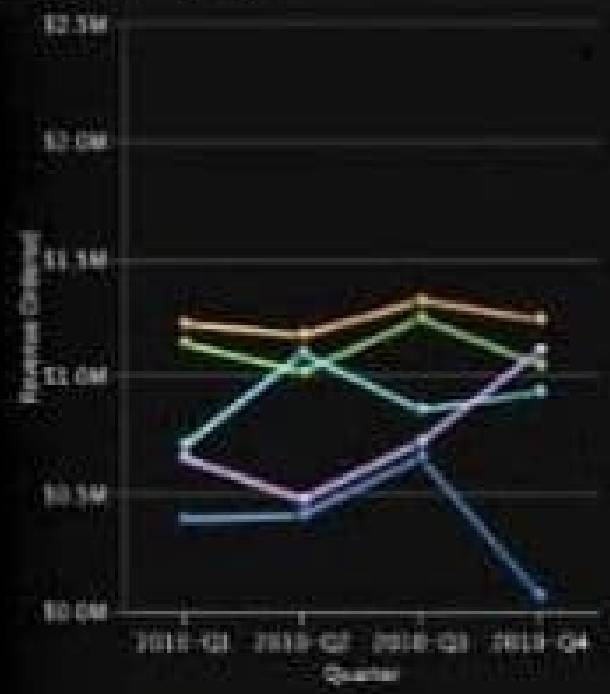
26.1%	$110 M	$882 M	$50 M	32%	22.3%
Market Share	Inquiry Revenue	Market Revenue	Order Revenue	Sales Growth	Win / Loss

REVENUE ORDERED

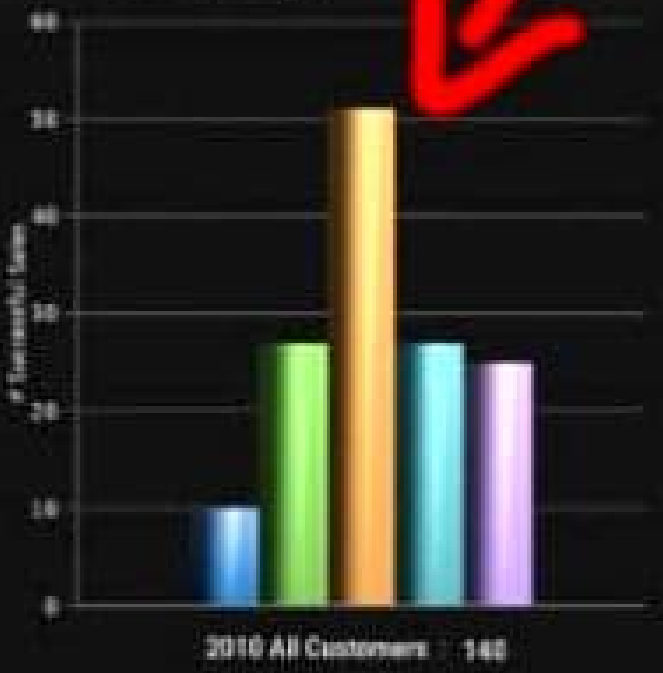
Alpha Revenue: $17,522,364

Alpha | Charger | Ram

REVENUE ORDERED - Alpha



SUCCESSFUL SALES - Alpha



2010 All Customers : 140

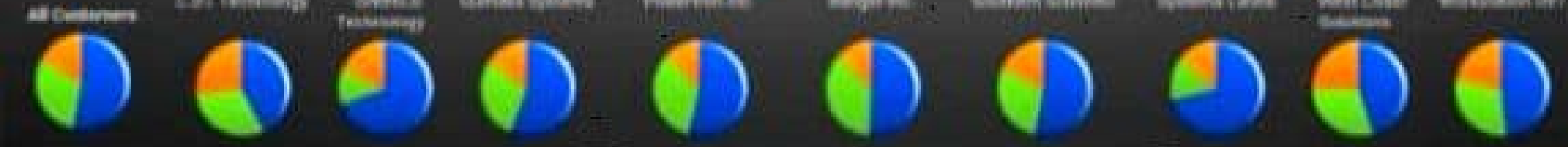
All Customers | C.E.I. Technology | Gulfstar Systems | Powertron Inc. | Ranger Inc. | Southern Scientific

Sales Dashboard

Introduction | Overview | Market Analysis | Product Shipments

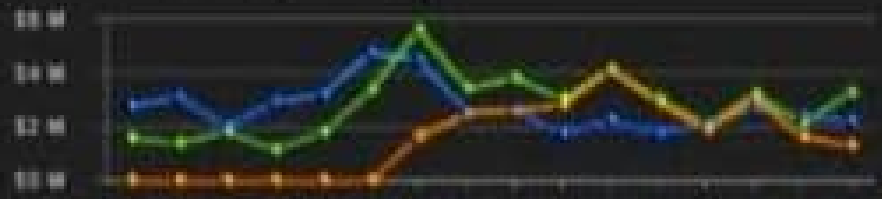
2010 ORDERED UNITS BY ALL CUSTOMERS

Alpha 11,429 Charger 8,573 Nova 3,619



CUSTOMER REVENUE BY STATE

ORDER REVENUE TREND (2007-2010)



MARKET REVENUE TREND (2007-2010)

CUSTOMER DETAILS

Customer	Inquiry Revenue	Order Revenue	Growth	List Price	Net Price	Order Units	Sales Growth %
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Summary	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]


Sales Dashboard
Sales Dashboard
Introduction
Overview
Market Analysis
Product Shipments
PRODUCT SHIPMENTS
Alpha
Product Type
Alpha Gold
Alpha Silver
Alpha Bronze
300
250
200
150
100
50
0
Shipment
2009 Qtr 1
2009 Qtr 2
2009 Qtr 3
2009 Qtr 4
2010 Qtr 1
2010 Qtr 2
2010 Qtr 3
2010 Qtr 4
Quarter

Product inventory overview by organization, with prior year comparative data