

Hash_Encoding_Technique

March 18, 2023

```
[1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

```
[2]: df = pd.read_csv('online_profit.csv')
```

```
[3]: df.head()
```

```
[3]:
```

	Marketing Spend	Administration	Transport	Area	Profit
0	114523.61	136897.80	471784.10	Dhaka	192261.83
1	NaN	151377.59	443898.53	Ctg	191792.06
2	153441.51	101145.55	407934.54	NaN	191050.39
3	144372.41	118671.85	383199.62	Dhaka	182901.99
4	142107.34	91391.77	366168.42	Rangpur	166187.94

```
[4]: df.isnull().sum()
```

```
[4]:
```

Marketing Spend	2
Administration	0
Transport	0
Area	3
Profit	1

dtype: int64

```
[5]: mean = df['Marketing Spend'].mean()
```

```
[6]: mean
```

```
[6]: 70691.35312500001
```

```
[7]: df['Marketing Spend'] = df['Marketing Spend'].fillna(mean)
```

```
[8]: df.head()
```

```
[8]: Marketing Spend Administration Transport Area Profit
0 114523.610000 136897.80 471784.10 Dhaka 192261.83
1 70691.353125 151377.59 443898.53 Ctg 191792.06
2 153441.510000 101145.55 407934.54 NaN 191050.39
3 144372.410000 118671.85 383199.62 Dhaka 182901.99
4 142107.340000 91391.77 366168.42 Rangpur 166187.94
```

```
[9]: df['Area'] = df['Area'].fillna(method='ffill')
```

```
[10]: median = df['Profit'].median()
```

```
[11]: median
```

```
[11]: 107404.34
```

```
[12]: df['Profit'] = df['Profit'].fillna(median)
```

```
[13]: df.head()
```

```
[13]: Marketing Spend Administration Transport Area Profit
0 114523.610000 136897.80 471784.10 Dhaka 192261.83
1 70691.353125 151377.59 443898.53 Ctg 191792.06
2 153441.510000 101145.55 407934.54 Ctg 191050.39
3 144372.410000 118671.85 383199.62 Dhaka 182901.99
4 142107.340000 91391.77 366168.42 Rangpur 166187.94
```

```
[14]: import category_encoders as ce
```

```
[15]: encoders = ce.HashingEncoder(cols='Area',n_components=3)
```

```
[16]: enco = encoders.fit_transform(df)
```

```
[17]: enco.head()
```

```
[17]: col_0 col_1 col_2 Marketing Spend Administration Transport Profit
0 0 1 0 114523.610000 136897.80 471784.10 192261.83
1 0 0 1 70691.353125 151377.59 443898.53 191792.06
2 0 0 1 153441.510000 101145.55 407934.54 191050.39
3 0 1 0 144372.410000 118671.85 383199.62 182901.99
4 1 0 0 142107.340000 91391.77 366168.42 166187.94
```

```
[18]: df = enco
```

```
[19]: df.head()
```

```
[19]: col_0 col_1 col_2 Marketing Spend Administration Transport Profit
0 0 1 0 114523.610000 136897.80 471784.10 192261.83
1 0 0 1 70691.353125 151377.59 443898.53 191792.06
```

2	0	0	1	153441.510000	101145.55	407934.54	191050.39
3	0	1	0	144372.410000	118671.85	383199.62	182901.99
4	1	0	0	142107.340000	91391.77	366168.42	166187.94

```
[20]: x = df.drop(['Profit'], axis=1)
```

```
[21]: y = df['Profit']
```

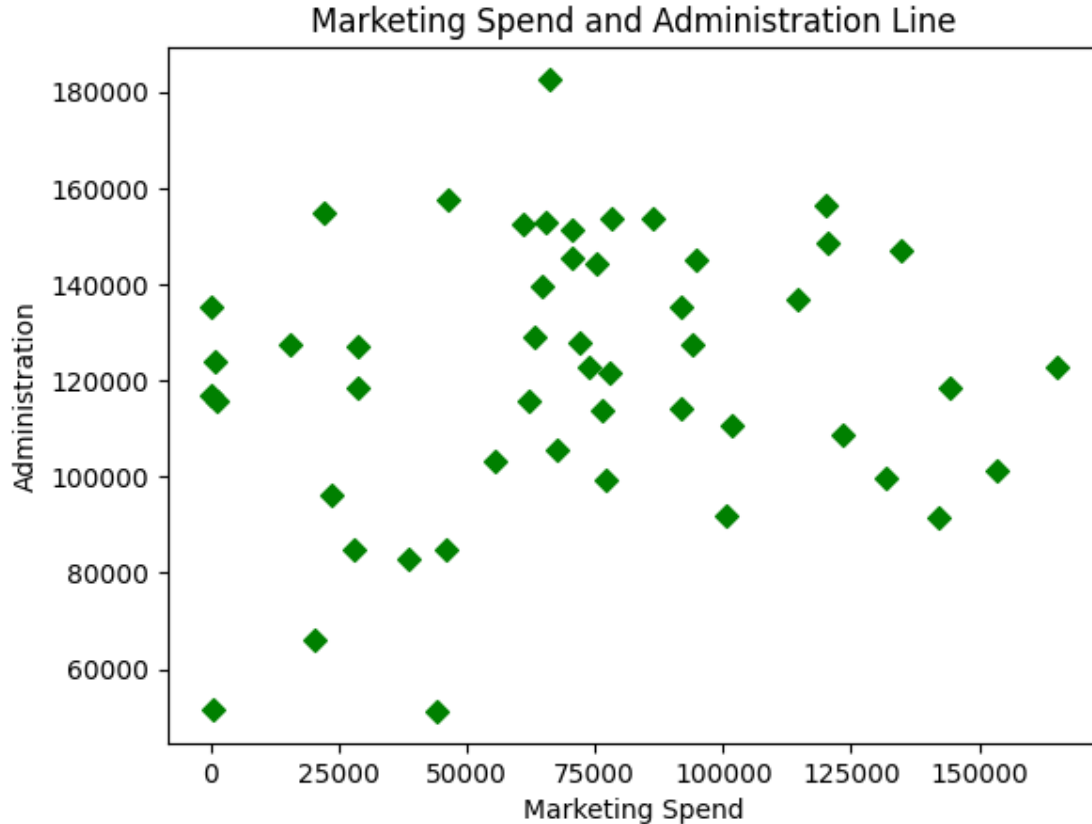
```
[22]: x.head()
```

```
[22]:
```

	col_0	col_1	col_2	Marketing Spend	Administration	Transport
0	0	1	0	114523.610000	136897.80	471784.10
1	0	0	1	70691.353125	151377.59	443898.53
2	0	0	1	153441.510000	101145.55	407934.54
3	0	1	0	144372.410000	118671.85	383199.62
4	1	0	0	142107.340000	91391.77	366168.42

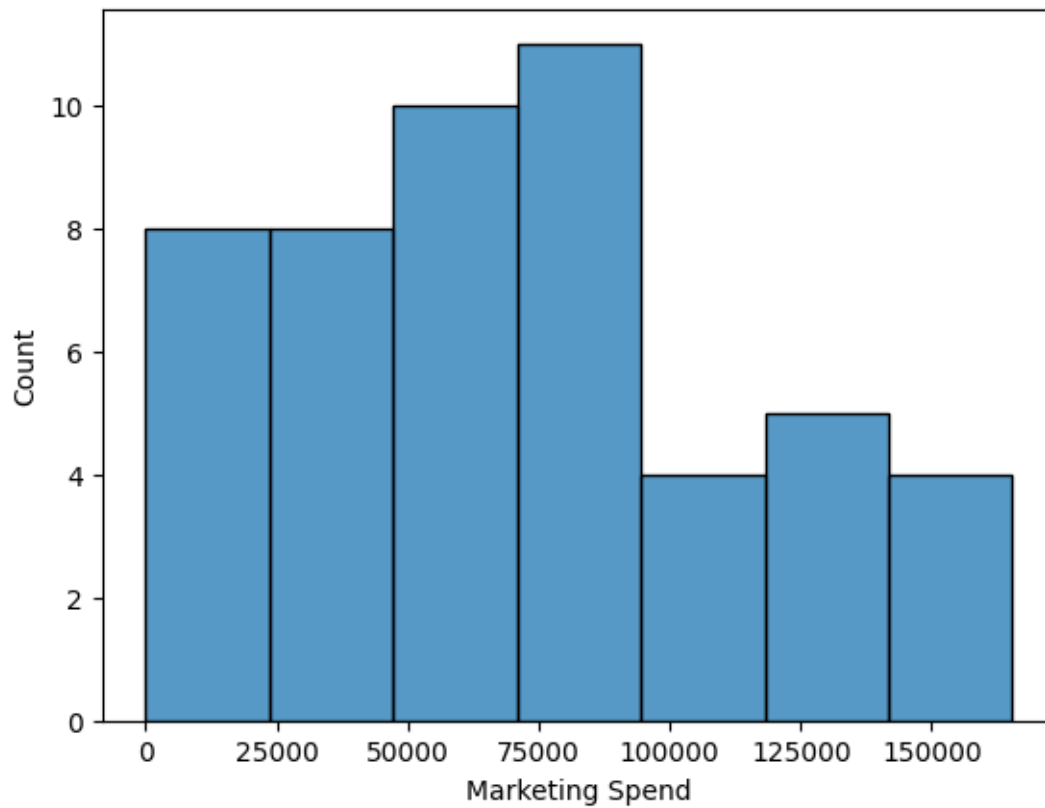
```
[23]: plt.title("Marketing Spend and Administration Line")
plt.xlabel("Marketing Spend")
plt.ylabel("Administration")
plt.scatter(df['Marketing Spend'],df['Administration'],marker="D",color="Green")
```

```
[23]: <matplotlib.collections.PathCollection at 0x173fd65c220>
```



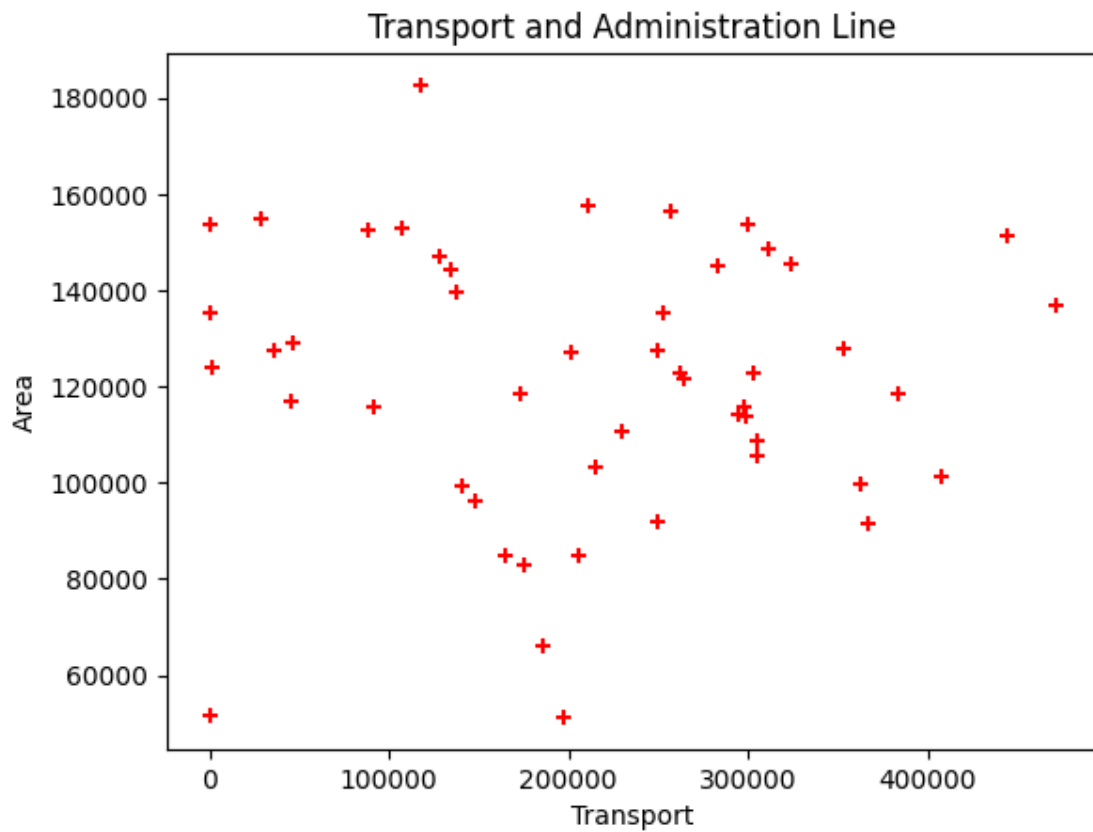
```
[24]: sns.histplot(df['Marketing Spend'])
```

```
[24]: <AxesSubplot: xlabel='Marketing Spend', ylabel='Count'>
```



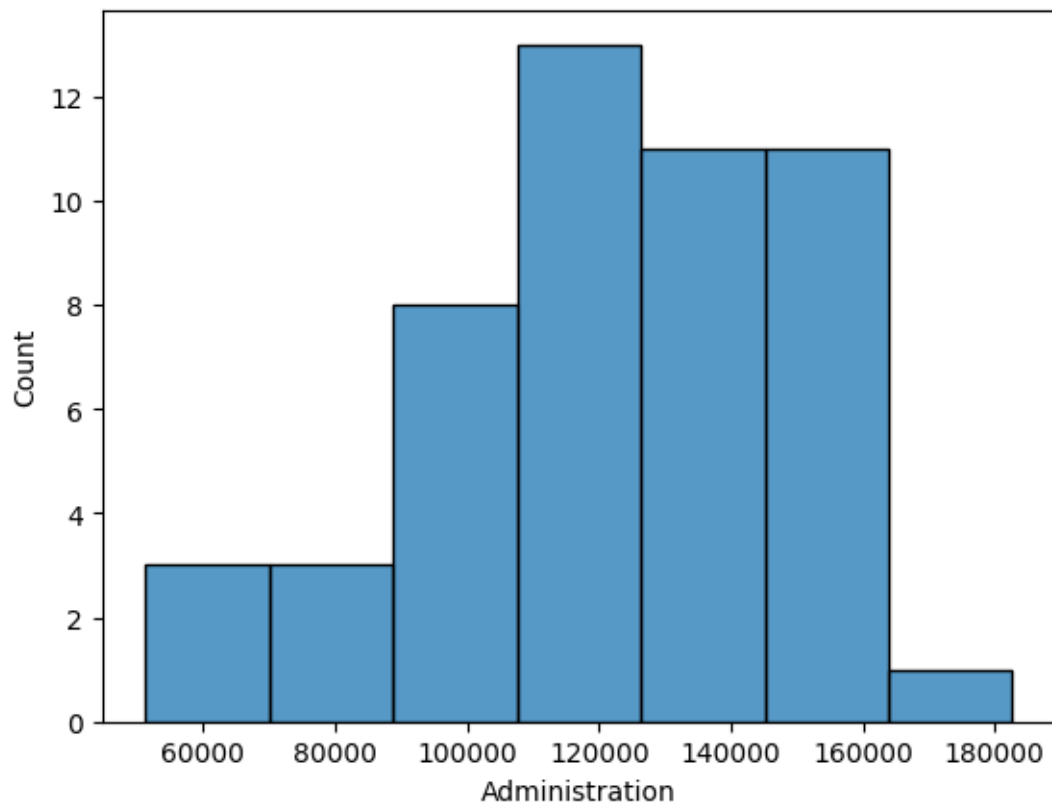
```
[25]: plt.title("Transport and Administration Line")
plt.xlabel("Transport")
plt.ylabel("Area")
plt.scatter(df['Transport'],df['Administration'],marker="+",color="Red")
```

```
[25]: <matplotlib.collections.PathCollection at 0x173f3537ac0>
```



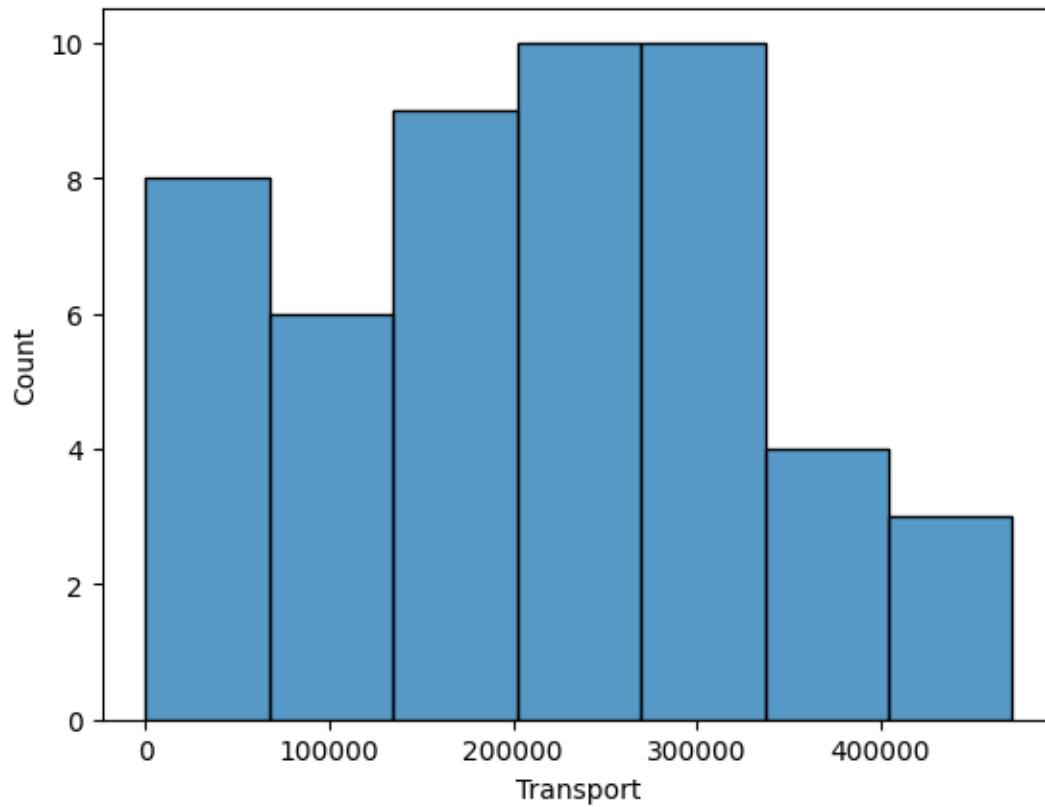
```
[26]: sns.histplot(df['Administration'])
```

```
[26]: <AxesSubplot: xlabel='Administration', ylabel='Count'>
```



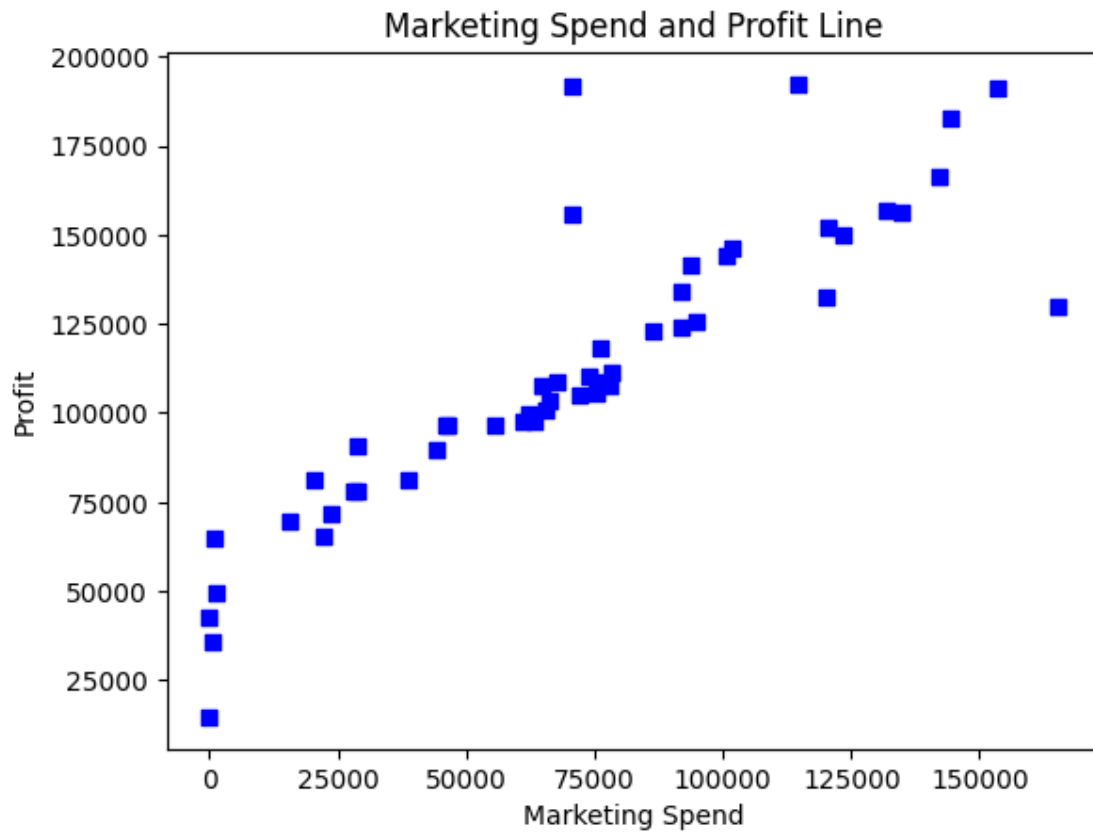
```
[27]: sns.histplot(df['Transport'])
```

```
[27]: <AxesSubplot: xlabel='Transport', ylabel='Count'>
```



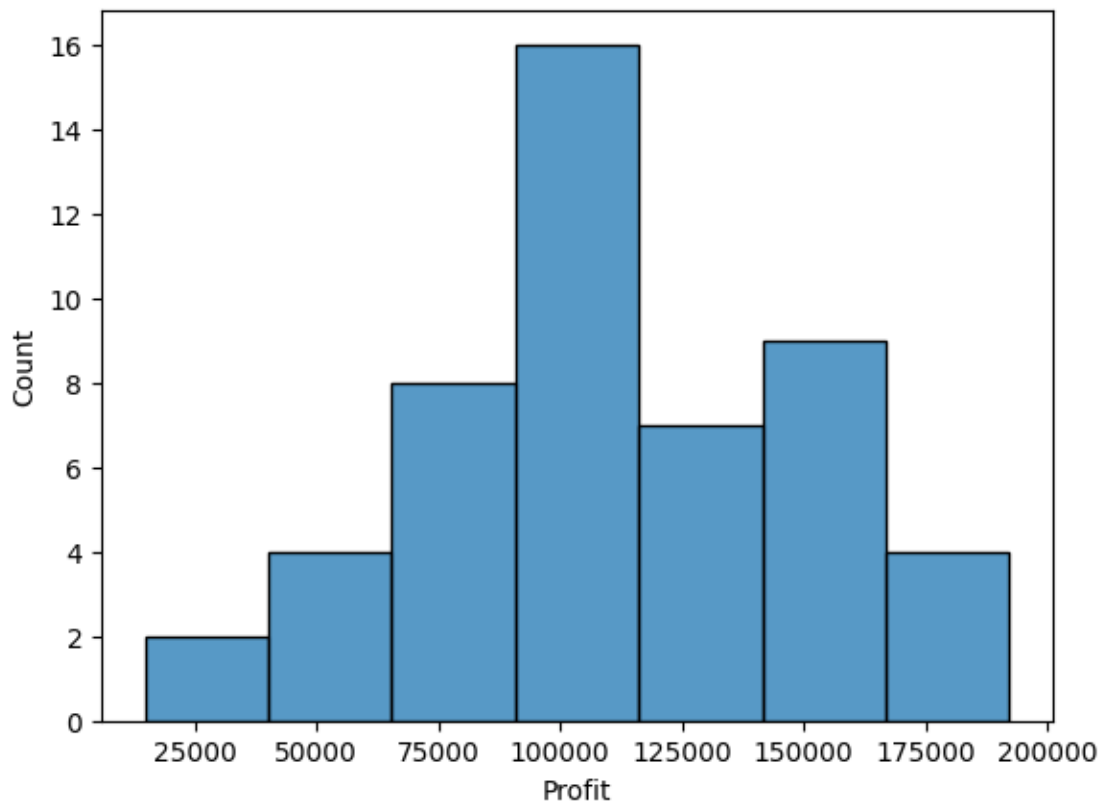
```
[28]: plt.title("Marketing Spend and Profit Line")
plt.xlabel("Marketing Spend")
plt.ylabel("Profit")
plt.scatter(df['Marketing Spend'],df['Profit'],marker="s",color="Blue")
```

```
[28]: <matplotlib.collections.PathCollection at 0x173f36ab5b0>
```



```
[29]: sns.histplot(df['Profit'])
```

```
[29]: <AxesSubplot: xlabel='Profit', ylabel='Count'>
```

```
[30]: from sklearn.model_selection import train_test_split
```

```
[31]: xtrain, xtest, ytrain, ytest = train_test_split(x,y,train_size=.  
    ↪70,random_state=42)
```

```
[32]: xtrain.shape
```

```
[32]: (35, 6)
```

```
[33]: xtest.shape
```

```
[33]: (15, 6)
```

```
[34]: ytrain.shape
```

```
[34]: (35,)
```

```
[35]: ytest.shape
```

```
[35]: (15,)
```

```
[36]: from sklearn.linear_model import LinearRegression
```

```
[37]: reg = LinearRegression()
```

```
[38]: reg.fit(xtrain,ytrain)
```

```
[38]: LinearRegression()
```

```
[39]: ytest
```

```
[39]: 13      134307.35
      39      81005.76
      30      99937.59
      45      64926.08
      17     125370.37
      48      35673.41
      26     105733.54
      25     107404.34
      32      97427.84
      19     122776.86
      12     141585.52
      4      166187.94
      37      89949.14
      8      152211.77
      3      182901.99
      Name: Profit, dtype: float64
```

```
[40]: reg.score(xtest.values,ytest)
```

```
[40]: 0.865858970562524
```

```
[41]: reg.coef_
```

```
[41]: array([-2.32540305e+03, -1.26402403e+03,  3.58942709e+03,  5.58987738e-01,
          1.65545425e-01,  1.52238111e-01])
```

```
[42]: reg.intercept_
```

```
[42]: 17127.450653436143
```

```
[43]: reg.predict([[0,1,0,114523.61,136897.80,471784.10]])
```

```
[43]: array([174367.04513447])
```