Ag10

December 28, 2023

[10]:
```
import pandas as pd
data = pd.read_csv('Data/Startups_Invest.csv')
data.head()
```

[10]:

	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

[11]:
```
data.info()
```

```
<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
```

[12]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

data['State'] = le.fit_transform(data['State'])
```

[13]:
```
from sklearn.model_selection import train_test_split

X = data.drop(['Profit'], axis=1)
y = data['Profit']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=44)
```

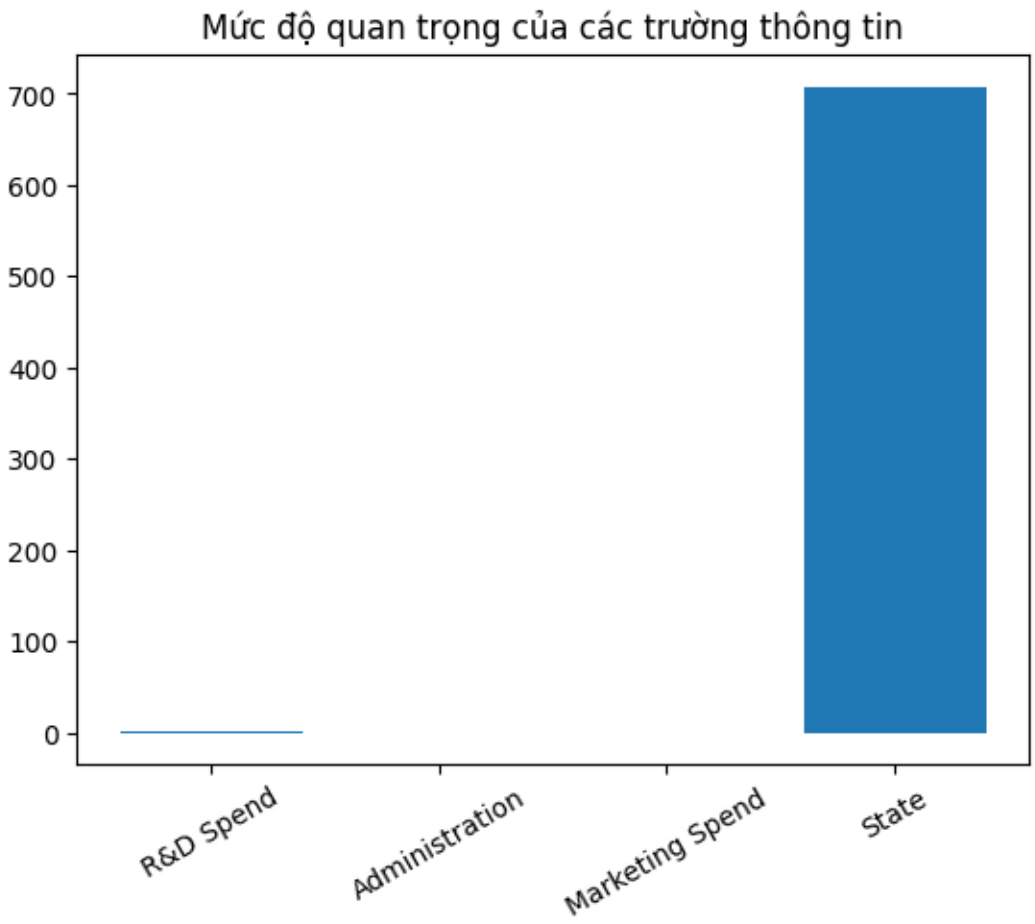
[14]:
```
print (X_train.shape)
print (X_test.shape)
```

```
(40, 4)
(10, 4)
```

[15]:
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error , r2_score

LR = LinearRegression()
LR.fit(X_train, y_train)
predictions_LR = LR.predict(X_test)

print ('MSE of LinearRegression= ', mean_squared_error(y_test, predictions_LR))
print ('R2_score of Linear Regression= ', r2_score(y_test, predictions_LR))
```

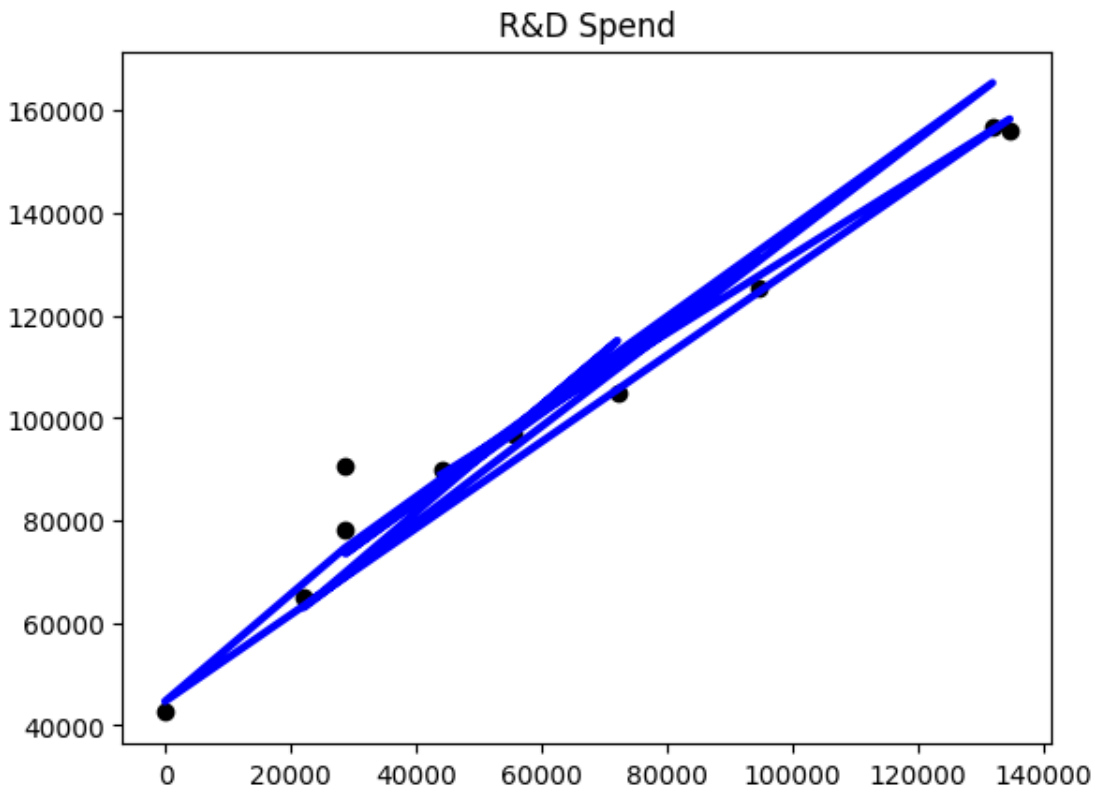
```
MSE of LinearRegression=  50354791.17254009
R2_score of Linear Regression=  0.9588715476383267
```

[16]:
```
import matplotlib.pyplot as plt
A = list(X.columns)
importance = LR.coef_

# plot feature importance
feature = data.columns
plt.bar(A, importance)
plt.title ('Mức độ quan trọng của các trường thông tin')
plt.xticks(rotation = 30)
plt.show()
```

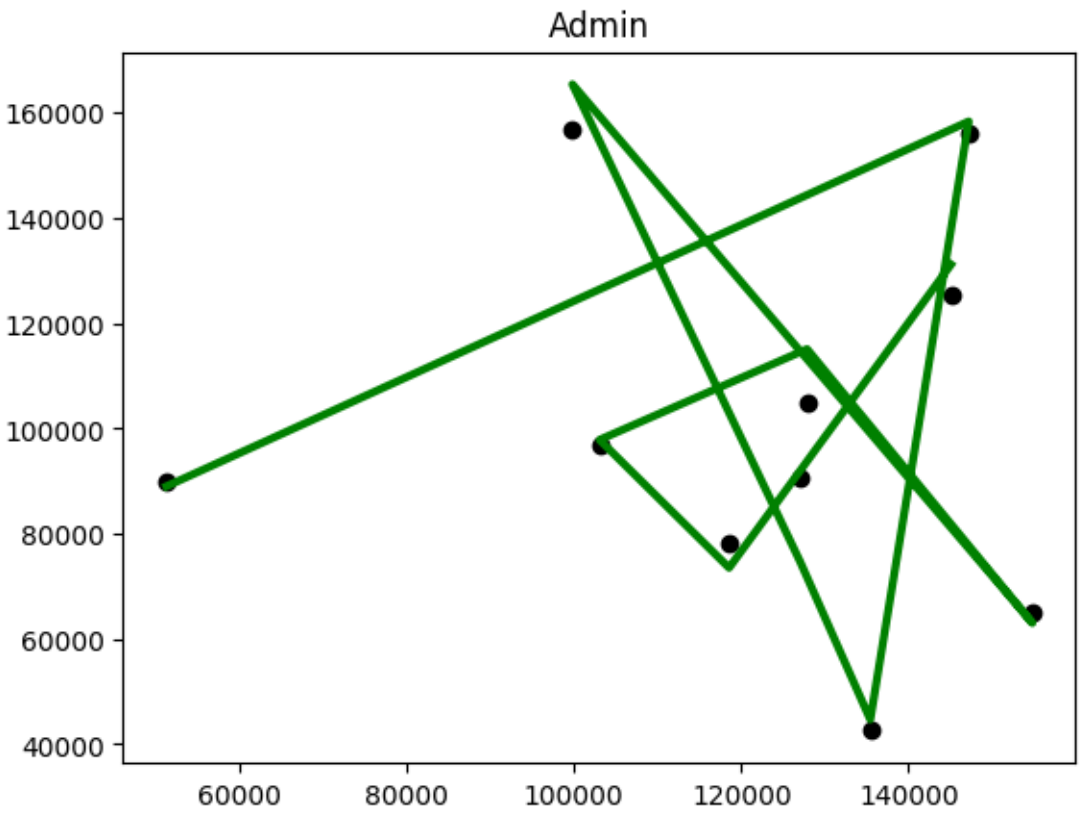


[17]:
```
X_RandD_test = X_test['R&D Spend']

X_Admin = X_test['Administration']

X_Marketing = X_test ['Marketing Spend']
```

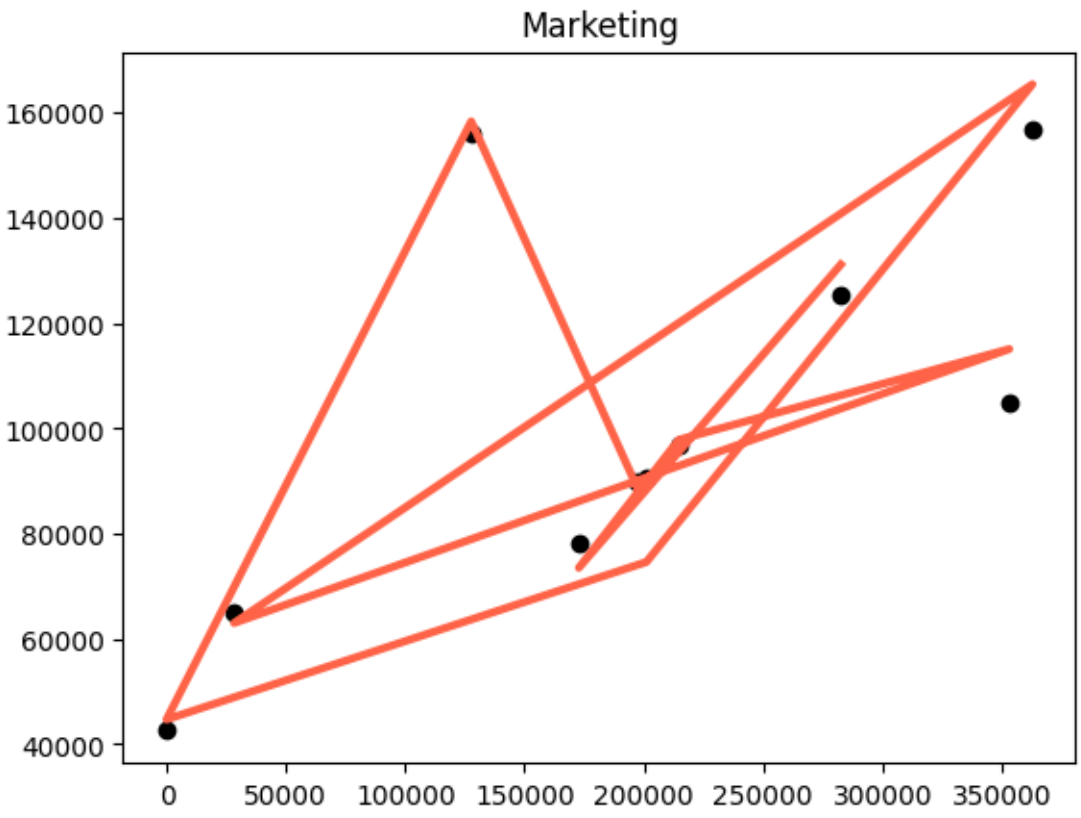
[18]:
```
plt.scatter(X_RandD_test, y_test, color="black")
plt.plot(X_RandD_test, predictions_LR, color="blue", linewidth=3)

plt.title ('R&D Spend')
plt.show()
```



[19]:
```
plt.scatter(X_Admin, y_test, color="black")
plt.plot(X_Admin, predictions_LR, color="green", linewidth=3)

plt.title ('Admin')
plt.show()
```



[20]:
```
plt.scatter(X_Marketing, y_test, color="black")
plt.plot(X_Marketing, predictions_LR, color="tomato", linewidth=3)

plt.title ('Marketing')
plt.show()
```


Marketing
160000
140000
120000
100000
80000
60000
40000
0
50000
100000
150000
200000
250000
300000
350000