

```

from sklearn.cluster import KMeans
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from matplotlib import pyplot as plt
%matplotlib inline

df=pd.read_csv("/content/drive/MyDrive/Datasets/50_Startups.csv")
df.head()
df.tail()

```

↗

	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

```

X = df.iloc[:,0:4]
y = df.iloc[:, -1]

```

```

df = pd.get_dummies(df)
df.head()

```

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

```

y=df['Profit']
X=df.drop(['Profit'],axis=1)

```

```
df.shape
```

```
(50, 7)
```

```

from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=0)

```

```
from sklearn.linear_model import LinearRegression
LR=LinearRegression().fit(X_train,y_train)
print(LR.score(X_train,y_train))
print(LR.score(X_test,y_test))
```

```
0.9511705268907698
0.9315815199152061
```

```
LR.predict(X_test)
```

```
array([104440.72582616, 132253.81567698, 132872.07174272, 71707.78771428,
       178678.9934985 , 115078.13124331, 66093.9297267 , 98759.7276409 ,
       114113.5994088 , 167979.49411476, 95786.77295823, 87785.3497979 ,
       110455.98228344])
```

```
X1=np.asarray(X)
y1=np.asarray(y)
md = LinearRegression()
md.fit(X1,y1)
print("R2 Score =",md.score(X1,y1))
```

```
R2 Score = 0.9507524843355148
```

```
import math
from sklearn.metrics import mean_squared_error
y_pred=LR.predict(X_test)
print(math.sqrt(mean_squared_error(y_test, y_pred)))
```

```
8591.234654128648
```