Aditi Rao 118A2088

Please find dataset 1 in the shared folder of Day 5. Apply simple linear regression on this dataset and estimate profit if money spent on R and D is 1,00,000 Rs. Upload py or ipynb file here. Write answer in Q.5

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```
from google.colab import files
uploaded = files.upload()
```

```
Choose Files  No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving data1.csv to data1.csv
```

```
dataset = pd.read_csv('data1.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 1].values
print(X)
print(y)
```

```
[[165349.2 ]
 [162597.7 ]
 [153441.51]
 [144372.41]
 [142107.34]
 [131876.9 ]
 [134615.46]
 [130298.13]
 [120542.52]
 [123334.88]
 [101913.08]
 [100671.96]
 [ 93863.75]
 [ 91992.39]
 [119943.24]
 [114523.61]
 [ 78013.11]
 [ 94657.16]
 [ 91749.16]
 [ 86419.7 ]
 [ 76253.86]
 [ 78389.47]
 [ 73994.56]
 [ 67532.53]
 [ 77044.01]
 [ 64664.71]
 [ 75328.87]
 [ 72107.6 ]
 [ 66051.52]
 [ 65605.48]
 [ 61994.48]
 [ 61136.38]
 [ 63408.86]
 [ 55493.95]
 [ 46426.07]
 [ 46014.02]
 [ 28663.76]
 [ 44069.95]
 [ 20229.59]
 [ 38558.51]
 [ 28754.33]
 [ 27892.92]
 [ 23640.93]
 [ 15505.73]
 [ 22177.74]
 [  1000.23]
 [  1315.46]
 [     0.  ]
```

```
      [   542.05]
      [     0.  ]]
     [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 ]
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, random_state = 0)
```

```
print(X_test)
print(y_test)
```

```
     [[ 66051.52]
      [100671.96]
      [101913.08]
      [ 27892.92]
      [153441.51]
      [ 72107.6 ]
      [ 20229.59]
      [ 61136.38]
      [ 73994.56]
      [142107.34]
      [ 55493.95]
      [ 46014.02]
      [ 75328.87]
      [ 46426.07]
      [ 91749.16]
      [130298.13]
      [119943.24]]
     [103282.38 144259.4  146121.95  77798.83 191050.39 105008.31  81229.06
       97483.56 110352.25 166187.94  96778.92  96479.51 105733.54  96712.8
      124266.9  155752.6  132602.65]
```

```
from sklearn.linear_model import LinearRegression
```

```
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results
y_pred = regressor.predict(X_test)
```

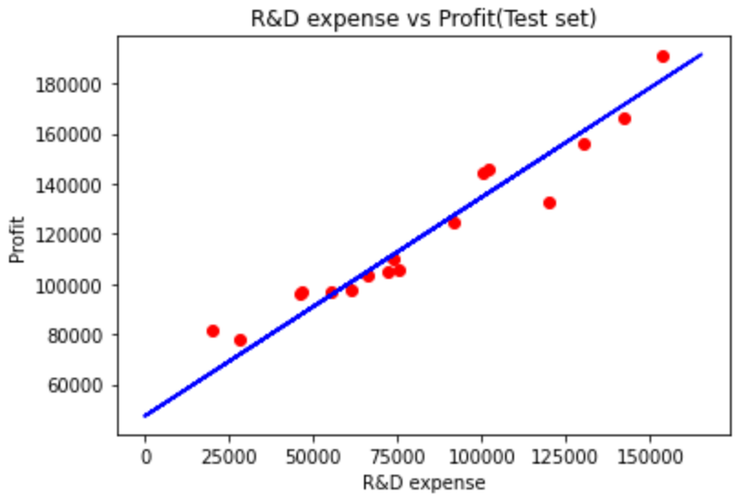
```
print(y_pred)
```

```
     [104951.99132775 135168.59853169 136251.84402288  71647.28960668
      181225.68979566 110237.71821498  64958.76015748 100662.0731542
      111884.65075337 171333.26315773  95737.37863888  87463.32662999
      113049.23217285  87822.96252712 127380.80777111 161026.22315596
      151988.50891056]
```

```
#visualizing training set results
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
```

```
[<matplotlib.lines.Line2D at 0x7f0258142c50>]
```


```
# Visualising the Test set results
plt.scatter(X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('R&D expense vs Profit(Test set)')
plt.xlabel('R&D expense')
plt.ylabel('Profit')
plt.show()
```




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

```
8665.191847945045
```

```
from statistics import mean
y_test_mean = mean(y_test)
rmse = rmse / y_test_mean
print(rmse)
```

```
0.07252631067599734
```

```
a = [[100000]]
sal = regressor.predict(a)
print(sal)
```

```
[134582.11403361]
```