МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО”
Факультет прикладної математики
Кафедра програмного забезпечення комп’ютерних систем

Лабораторна робота №4

з дисципліни: «Технології оброблення великих даних»

на тему: «Побудова лінійної регресії в Python»

Виконав
студент III курсу каф. ПЗКС ФПМ
групи КП-82
Мельничук Олексій Геннадійович

Перевірила
доц. каф. ПЗКС ФПМ
Олещенко Л.М.

Київ 2021

1. Індивідуальне завдання

Мета: ознайомитись з поняттями простої лінійної регресії та роботи з наданими даними для прогнозування в Python.

У статистиці лінійна регресія - це спосіб моделювання взаємозв'язку між залежною змінною p і незалежною змінною x. У цій лабораторній роботі ви проаналізуєте дані про продажі району та побудуєте просту лінійну регресію для прогнозування річного чистого обсягу продажів на основі кількості магазинів у районі.

2. Хід роботи

Stores_dist.txt

```
Gender,FSIQ,VIQ,PIQ,Weight,Height,MRI_Count
Female,133,132,124,118,64.5,816932
Male,140,150,124,NaN,72.5,1001121
Male,139,123,150,143,73.3,1038437
Male,133,129,128,172,68.8,965353
Female,137,132,134,147,65.0,951545
Female,99,90,110,146,69.0,928799
Female,138,136,131,138,64.5,991305
Female,92,90,98,175,66.0,854258
Male,89,93,84,134,66.3,904858
Male,133,114,147,172,68.8,955466
Female,132,129,124,118,64.5,833868
Male,141,150,128,151,70.0,1079549
Male,135,129,124,155,69.0,924059
Female,140,120,147,155,70.5,856472
Female,96,100,90,146,66.0,878897
Female,83,71,96,135,68.0,865363
Female,132,132,120,127,68.5,852244
Male,100,96,102,178,73.5,945088
Female,101,112,84,136,66.3,808020
Male,80,77,86,180,70.0,889083
Male,83,83,86,NaN,NaN,892420
Male,97,107,84,186,76.5,905940
Female,135,129,134,122,62.0,790619
Male,139,145,128,132,68.0,955003
Female,91,86,102,114,63.0,831772
Male,141,145,131,171,72.0,935494
Female,85,90,84,140,68.0,798612
Male,103,96,110,187,77.0,1062462
Female,77,83,72,106,63.0,793549
Female,130,126,124,159,66.5,866662
Female,133,126,132,127,62.5,857782
Male,144,145,137,191,67.0,949589
Male,103,96,110,192,75.5,997925
Male,90,96,86,181,69.0,879987
Female,83,90,81,143,66.5,834344
Female,133,129,128,153,66.5,948066
Male,140,150,124,144,70.5,949395
Female,88,86,94,139,64.5,893983
Male,81,90,74,148,74.0,930016
Male,89,91,89,179,75.5,935863
```

Частина 1: Імпорт бібліотек та даних

Крок 2: Імпортуйте дані.

```
#step2
salesDist = pd.read_csv('./stores-dist.csv')
print(salesDist.head())
```

```
C:\Users\datru\Desktop\study2021\big_data\lab4>python3 main.py
   district  annual net sales  number of stores in district
0         1             231.0                            12
1         2             156.0                            13
2         3              10.0                            16
3         4             519.0                             2
4         5             437.0                             6

C:\Users\datru\Desktop\study2021\big_data\lab4>
```

Заголовки імпортовано правильно

```
20    salesDist = salesDist.rename(columns={'annual net
      sales':'sales','number of stores in district':'stores'})
21    print(salesDist.head())
```

```
C:\Users\datru\Desktop\study2021\big_data\lab4>python3 main.py
   district  sales  stores
0         1  231.0      12
1         2  156.0      13
2         3   10.0      16
3         4  519.0       2
4         5  437.0       6

C:\Users\datru\Desktop\study2021\big_data\lab4>
```

Частина 2: Складання даних

Крок 1: Визначте співвідношення

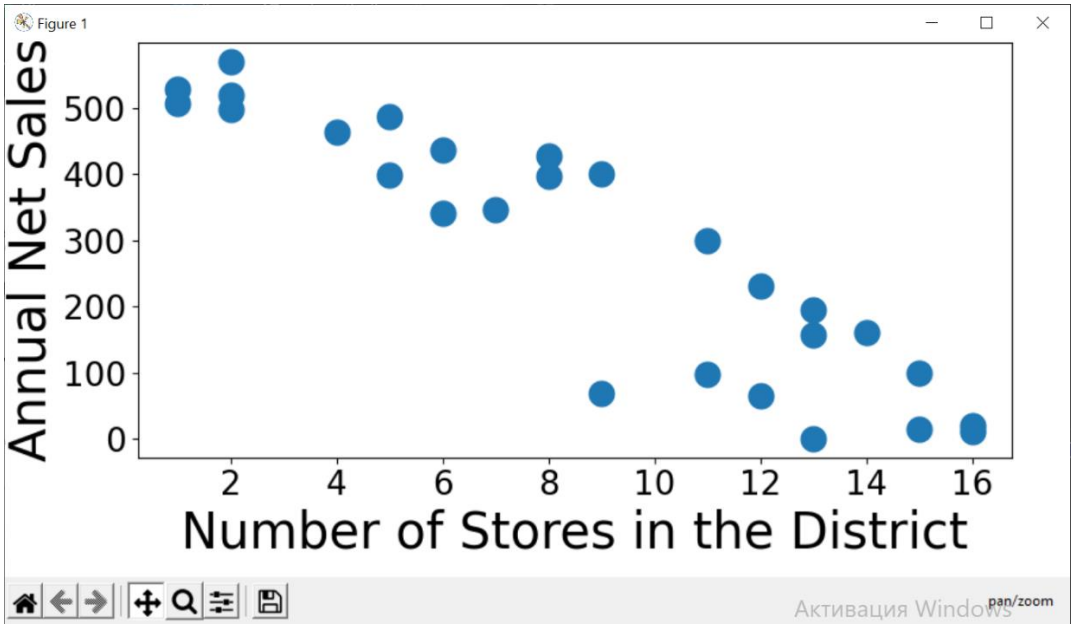
```
22    #PART 2
23
24    #step 1
25    sales = salesDist.drop(columns={'district'})
26    print(sales.head())
```

```
C:\Users\datru\Desktop\study2021\big_data\lab4>python3 main.py
   annual net sales  number of stores in district
0             231.0                           12
1             156.0                           13
2              10.0                           16
3             519.0                            2
4             437.0                            6

C:\Users\datru\Desktop\study2021\big_data\lab4>
```

Кореляція є від’ємною

Крок 2: Створіть сюжет.

```
28  #step2
29  y = sales['sales']
30  x = sales.stores
31
32  plt.figure(figsize=(20,10))
33
34  plt.plot(x,y, 'o', markersize = 15)
35
36  plt.ylabel('Annual Net Sales', fontsize = 30)
37  plt.xlabel('Number of Stores in the District', fontsize =
    30)
38
39  plt.xticks(fontsize = 20)
40  plt.yticks(fontsize = 20)
41
42  plt.show()
```



Частина 3: Побудуйте просту лінійну регресію

Крок 1: Обчисліть нахил та перетин Y-лінії лінійної регресії.

```
45  # #PART 3
46  # #step1
47  m, b = np.polyfit(x,y,1)
48  print ('The slope of line is {:.2f}.'.format(m))
49  print ('The y-intercept is {:.2f}.'.format(b))
50  print ('The best fit simple linear regression line is {:2f}
    x + {:.2f}.'.format(m, b))
```

```
C:\Users\datru\Desktop\study2021\big_data\lab4>python3 main.py
   sales  stores
0  231.0      12
1  156.0      13
2   10.0      16
3  519.0       2
4  437.0       6
The slope of line is -35.79.
The y-intercept is 599.38.
The best fit simple linear regression line is -35.787085x + 599.38.

C:\Users\datru\Desktop\study2021\big_data\lab4>
```

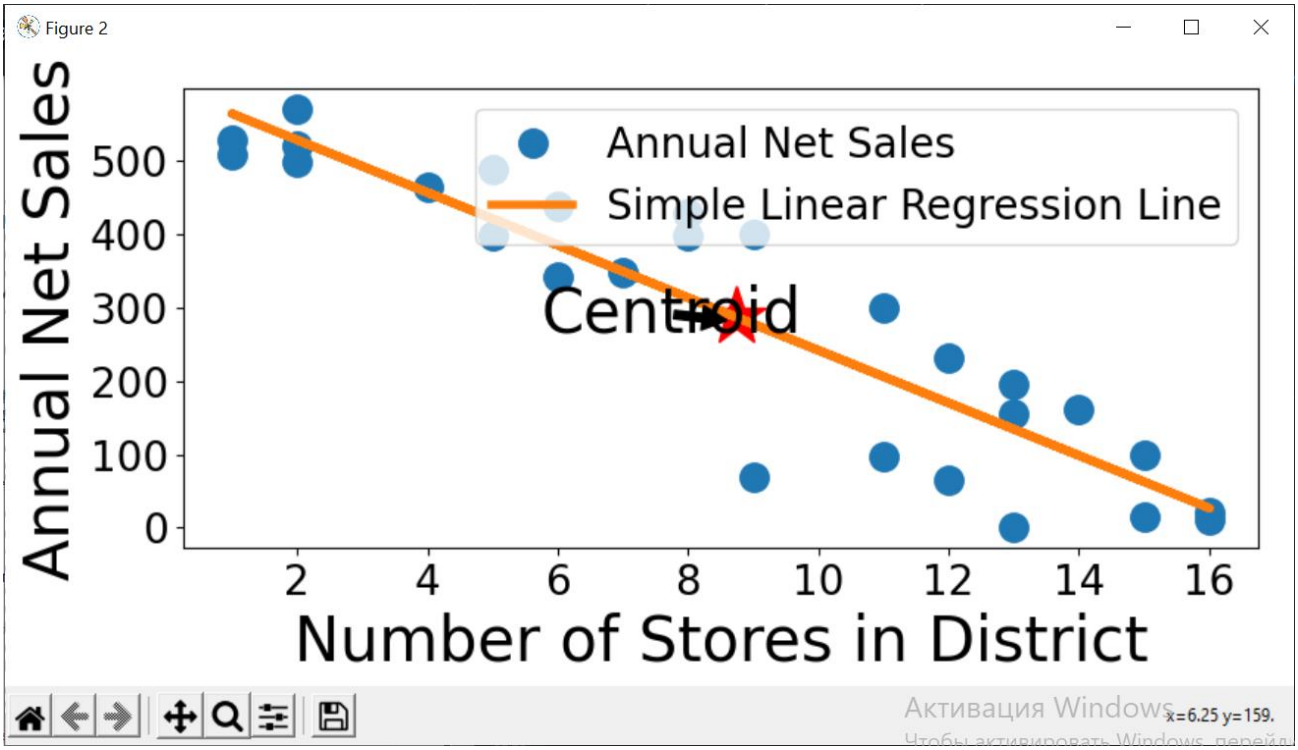
Крок 2: Обчисліть центроїд

```
52  y_mean = y.mean()
53  x_mean = x.mean()
54  print ('The centroid for this dataset is x = {:.2f} and y
    = {:.2f}.'.format(x_mean, y_mean))
```

```
C:\Users\datru\Desktop\study2021\big_data\lab4>python3 main.py
   sales  stores
0  231.0      12
1  156.0      13
2   10.0      16
3  519.0       2
4  437.0       6
The centroid for this dataset is x = 8.74 and y = 286.57.
```

Крок 3: Накладіть лінію регресії та центральну точку на графіку

```python
#step3
plt.figure(figsize=(20,10))

plt.plot(x,y, 'o', markersize = 14, label = "Annual Net Sales")

plt.plot(x_mean,y_mean, '*', markersize = 30, color = "r")

plt.plot(x, m*x + b, '-', label = 'Simple Linear Regression Line', linewidth = 4)

plt.ylabel('Annual Net Sales', fontsize = 30)
plt.xlabel('Number of Stores in District', fontsize = 30)

plt.xticks(fontsize = 20)
plt.yticks(fontsize = 20)

plt.annotate('Centroid', xy=(x_mean-0.1, y_mean-5), xytext=(x_mean-3, y_mean-20),
arrowprops=dict(facecolor='black', shrink=0.05), fontsize = 30)

plt.legend(loc = 'upper right', fontsize = 20)

plt.show()
```



Крок 4: Прогнозування

```
78  #step4
79  def predict(query):
80      if query >= 1:
81          predict = m * query + b
82          return predict
83      else:
84          print ("You must have >= 1 store in the district to predict the annual net sales.")
85
86  print(predict(4))
```

Прогнозований чистий продаж з 4-х магазинів:

```
C:\Users\datru\Desktop\study2021\big_data\lab4>python3 main.py
456.2313681207654
```

Висновки

В ході виконанні лабораторної роботи були використані навички моделювання лінійної регресії на базі набору даних з обсягів продажів в районах з допомогою бібліотек для мови Python, а саме Numpy, Pandas, Matplotlib. Ці бібліотеки були використані для зручної побудови графіків регресії чистого обсягу даних.