LAPORAN
PRAKTIKUM BIG DATA ANALYTIC
Pertemuan Ke - 3

Dosen :
Sri Redjeki, S.Si., M.Kom.

Disusun oleh :
RAHADIYAN BONDAN PERMADI
215411119

Universitas Teknologi Digital Indonesia
UTDI
YOGYAKARTA
2022

Dasar Teori

Sama seperti Bahasa pemrograman lain, Python juga memiliki banyak library yang dapat digunakan untuk membantu kita dalam membangun sebuah aplikasi. Dalam praktikum ini, fungsi yang akan digunakan adalah Matplotlib, Pandas, Numpy.

Matplotlib Merupakan library yang paling sering digunakan oleh data science karena dapat digunakan untuk memvisualisasikan data (misalnya dalam bentuk grafis). Matplotlib memiliki Plot untuk menampilkan data secara 2D atau 3D. Plot sendiri dapat berupa garis, sebaran, histogram.

Pandas (Python Data Analysis) merupakan Library yang dapat digunakan untuk manipulasi dan analisis data yang memiliki struktur data yang diperlukan untuk membersihkan data mentah ke dalam sebuah bentuk yang cocok untuk analisis (yaitu tabel). Pandas dapat menyelaraskan data untuk perbandingan dan penggabungan dataset, penanganan data yang hilang, dll. Struktur data dasar pandas dinamakan DataFrame, yaitu sebuah koleksi kolom berurutan dengan nama dan jenis, dengan demikian merupakan 2 sebuah tabel yang tampak seperti database dimana sebuah baris tunggal mewakili sebuah contoh tunggal dan kolom mewakili atribut tertentu. Dengan adanya fitur DataFrame memudahkan untuk membaca sebuah file dan menjadikannya tabel, kita juga dapat mengolah suatu data dengan menggunakan operasi seperti join, distinct, group by, agregasi, dan teknik lainnya yang terdapat pada SQL. Banyak format file yang dapat dibaca menggunakan Pandas, seperti file .txt, .csv, .tsv dan lainnya.

Numpy (Numeric Python) package Python yang digunakan sebagai alternative List Python, yaitu Numpy array (mirip dengan List). NumPy biasanya digunakan bersamaan dengan package lain seperti Matplotlib dan SciPy. Library ini memungkinkan kita bekerja dengan matriks dan array multidimensi yang besar. Selain itu, NumPy juga menyediakan fungsi tingkat tinggi untuk melakukan operasi matematika.

Kebutuhan Alat

1. Python (Anaconda / Miniconda)

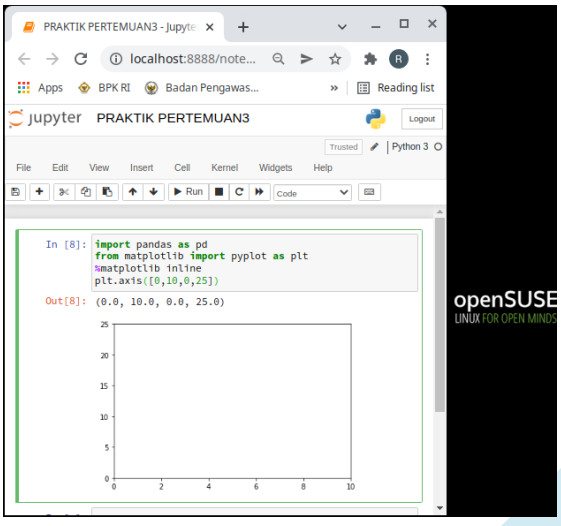
2. Jupyter Notebook

Langkah – Langkah dalam praktikum

Praktikum :

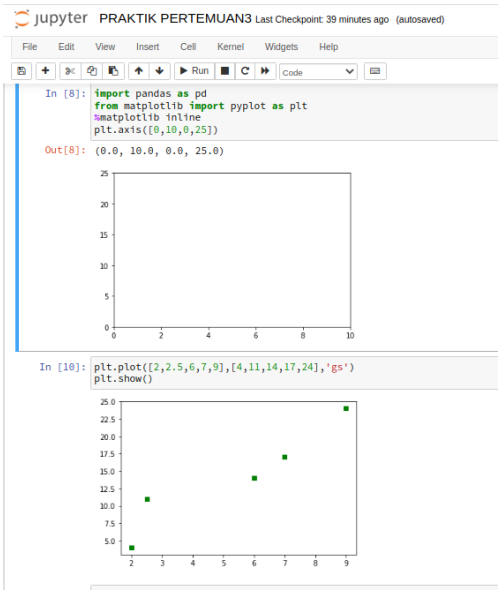
1. Menggunakan fungsi Membuat Ploting data pada sumbu ordinat a. Membuat garis ordinat untuk ploting data menggunakan perintah dibawah ini :

```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
plt.axis([0,10,0,25])
```



b. Menampilkan data ordinat untuk data {(2,4), (2.5,11), (6,14), (7,17), (9,24)} gunakan coding dibawah ini

```
plt.plot([2,2.5,6,7,9],[4,11,14,17,24],'gs')
plt.show()
```

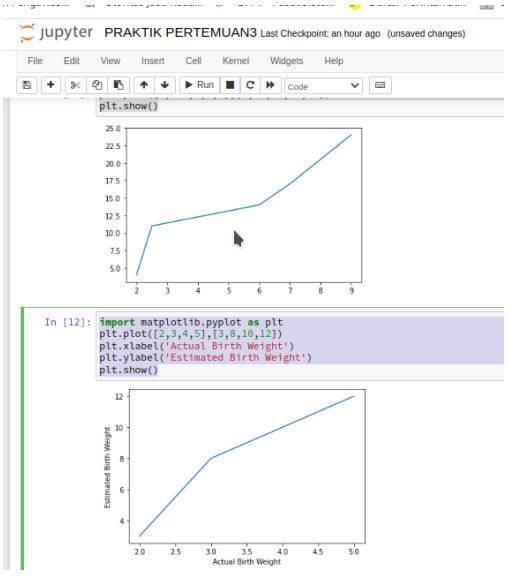


c. Membuat garis pada titik-titik ordinat diatas dengan menghilangkan “gs” dan tampilan dibawah ini :

```
plt.plot([2,2.5,6,7,9],[4,11,14,17,24])
plt.show()
```

d. Membuat ploting data dengan memberi nama variabel sumbu x dan sumbu y, seperti dibawah ini :

```
import matplotlib.pyplot as plt
plt.plot([2,3,4,5],[3,8,10,12])
plt.xlabel('Actual Birth Weight')
plt.ylabel('Estimated Birth Weight')
plt.show()
```



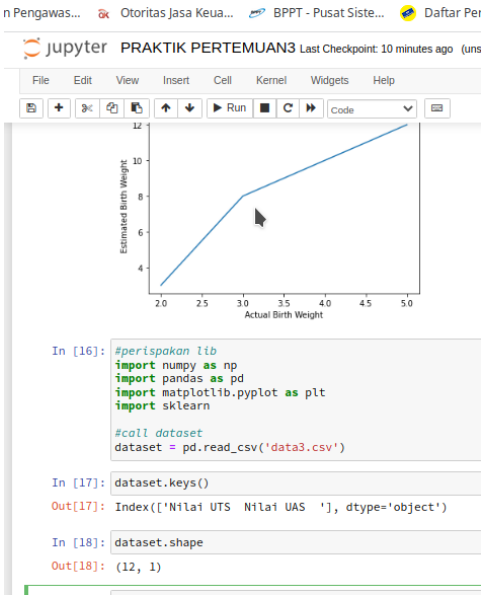
2. Manajemen data file csv

a. Menampilkan data → Buatlah data seperti dibawah ini dan simpan dengan nama data3.csv

Untuk memastikan data tersimpan dan dapat dibaca oleh python gunakan script dibawah ini :

Input [17] digunakan untuk menampilkan nama variabel pada data3.csv

Input [18] digunakan untuk melihat jumlah record dan jumlah atribut dari data3.csv



b. Scatter Plot Data

```
import pandas as pd
pd.plotting.register_matplotlib_coverters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
insurance = 'insurance.csv'
```

```
In [ ]:
import pandas as pd
pd.plotting.register_matplotlib_coverters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

```
In [27]:
insurance = 'insurance.csv'
insurance_data = pd.read_csv(insurance)
insurance_data.head()
```

Out[27]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

insurance_data = pd.read_csv(insurance)

insurance_data.head()

dapat kita tampilkan semua data menggunakan perintah

insurance = pd.read_csv("insurance.csv")

print dataframe

insurance

Terdapat 1337 Baris data dan 7 Kolom data.

```
In [30]: insurance = pd.read_csv("insurance.csv")
         # print dataframe
         insurance
```

Out[30]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

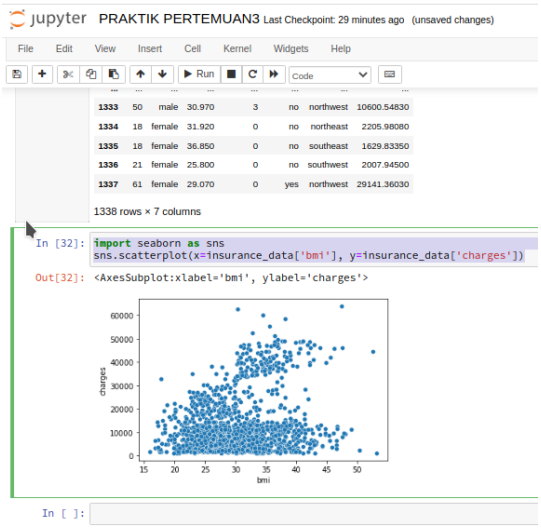
1338 rows × 7 columns

```
In [ ]:
```

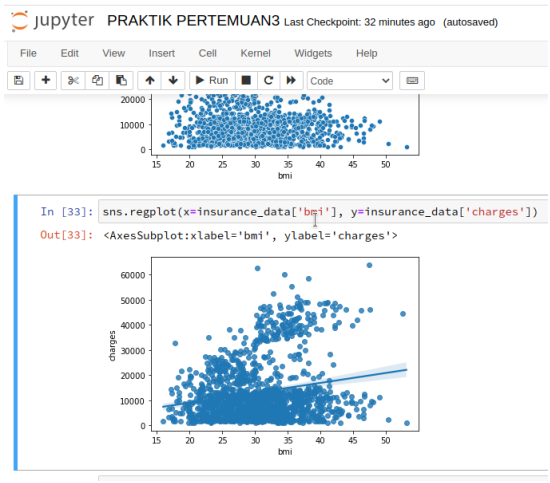
Fungsi untuk membuat scatter plot, jangan lupa import dulu lib nya seaborn :)

import seaborn as sns

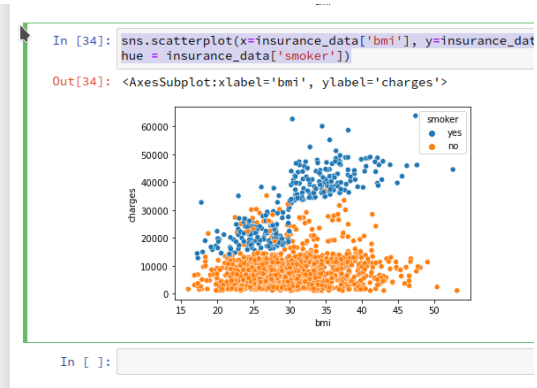
sns.scatterplot(x=insurance_data['bmi'], y=insurance_data['charges'])



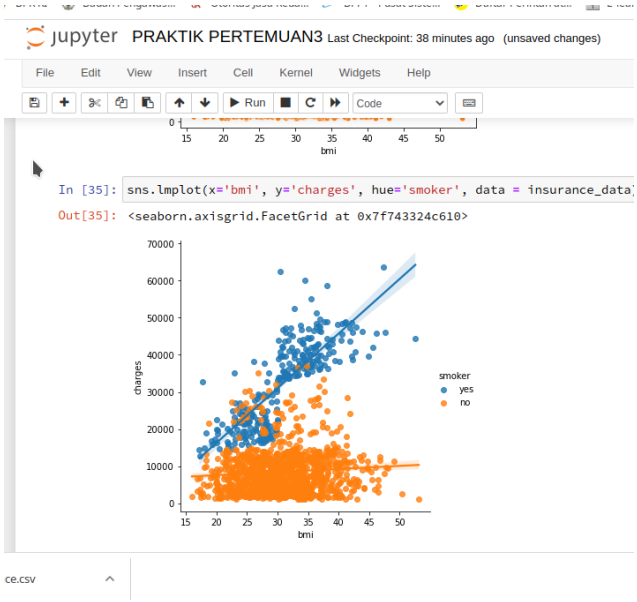
Untuk membuat plotting regresi dengan variabel x = bmi dan y = charges



Menampilkan plot data kategori dari salah satu variabel smoker



Menampilkan prediksi data tentang smoker melalui garis regresi



LATIHAN DAN TUGAS

1. Jelaskan fungsi Seaborn.

Seaborn merupakan library python yang berfungsi untuk memvisualisasikan data secarastatistik agar data terhliat menarik

2. Jelaskan perbedaan scatterplot, regplot, lmplot

scatterplot memvisualisasikan data menggunakan titik – titik yang terletak diantara 2 sumbu, sedangkan **regplot** berfungsi membuat untuk plotting regresi dengan variabel dua sumbu x dan y, dan **lmplot** berfungsi untuk menampilkdan data prediksi melalui garis regresi.

Demikian laporan dan tugas Pertemuan Ketiga yang dapat saya rangkum dan saya kerjakan, saya dapat mempraktekkan penggunaan python dan mampu memahami penggunaan numpy, matplotlib, scatterplot, regplo dan lmplot.

==========*Terimakasih*==========