MODUL

BIG DATA ANALYTIC



Disusun oleh :

Sri Redjeki, S.Si, M.Kom.

Maria Mediatrix S., S.Kom., M.Eng.

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

AKAKOM

YOGYAKARTA

2020

# KATA PENGANTAR

# DAFTAR ISI

[KATA PENGANTAR ii](#_Toc15905388)

[DAFTAR ISI iii](#_Toc15905389)

[MODUL 1 NAMA POKOK BAHASAN 1](#_Toc15905390)

# MODUL 3 EKSPLORASI DATA

|  |  |
| --- | --- |
|  | CAPAIAN PEMBELAJARAN |
| 1. Mampu menampilkan data menggunakan beberapa library di Python | |

|  |  |
| --- | --- |
|  | KEBUTUHAN ALAT/BAHAN/SOFTWARE |
| 1. Python (Anaconda / Miniconda) 2. Jupyter Notebook | |

|  |  |
| --- | --- |
|  | 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

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 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.

Untuk membuka file csv dengan perintah read\_excel(nama\_file\_excel) maka data akan masuk ke dataFrame pandas. Hasil dataFrame dapat ditampilkan dengan print. Beberapa fungsi lain dari pandas:

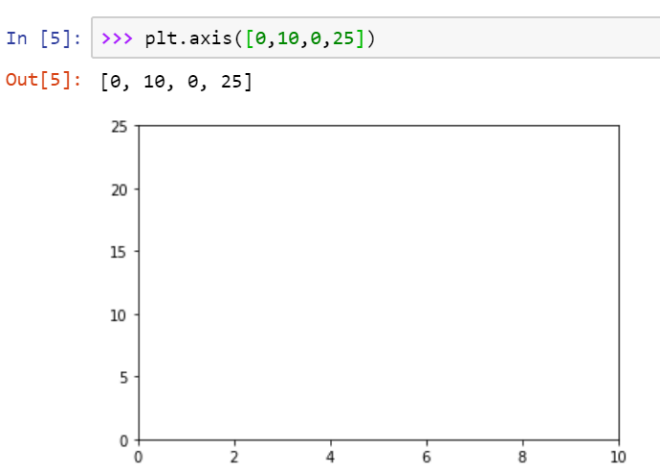
* df.head() 🡪 pengambilan 5 data pertama
* df.head(10) 🡪 pengambilan 10 data pertama
* df.tail(8) 🡪 pengambilan 8 data terakhir
* df.columns 🡪 menampilkan nama kolom
* df[[„nama', „alamat']] 🡪 pilih kolom tertentu
* df[“field”].min() 🡪 data terkecil
* df[“field”].max) 🡪 data terbesar
* df[“field”].mean() 🡪 rata-rata
* df[“field”].sum() 🡪 jumlah
* df[“field”].count() 🡪 cacah
* df[df['field'] operator\_perbandingan value] 🡪 filter data, misal df[df[„nilai‟] > 50]

Numpy

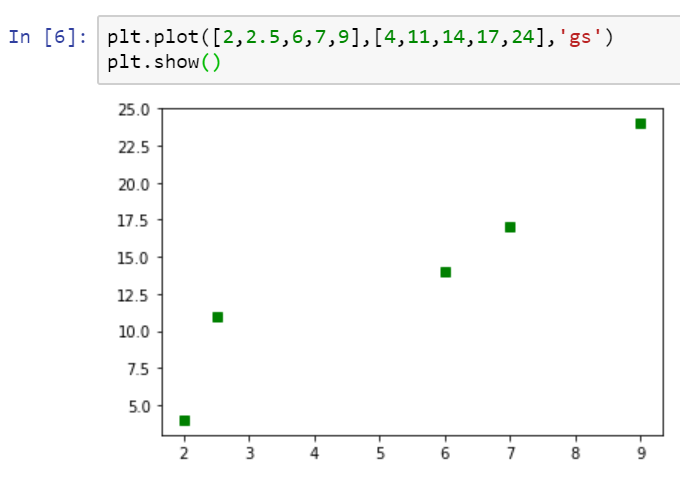
Numeric Python (NumPy) package Python yang digunakan sebagai alternative List Python, yaitu Numpy array (mirip dengan List). NumPy biasanya digunakan bersamaan dengan package lain seperti Matplotlib dan SciPy. Library ini memungkinkan kita bekerja dengan matriks dan array multidimensi yang besar. Selain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

|  |  |
| --- | --- |
|  | PRAKTIK |

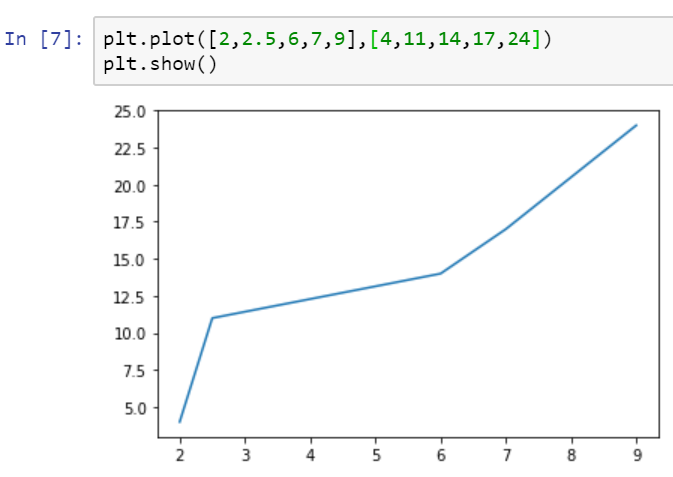
1. Menggunakan fungsi Membuat Ploting data pada sumbu ordinat
2. Membuat garis ordinat untuk ploting data menggunakan perintah dibawah ini :



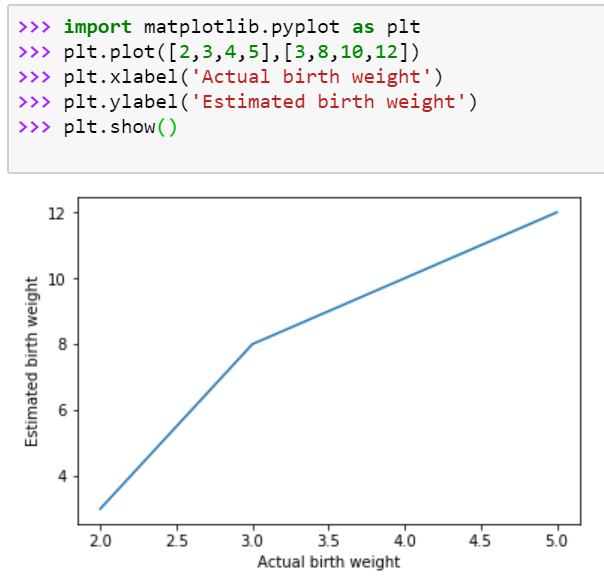
1. Menampilkan data ordinat untuk data {(2,4), (2.5,11), (6,14), (7,17), (9,24)} gunakan coding dibawah ini



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



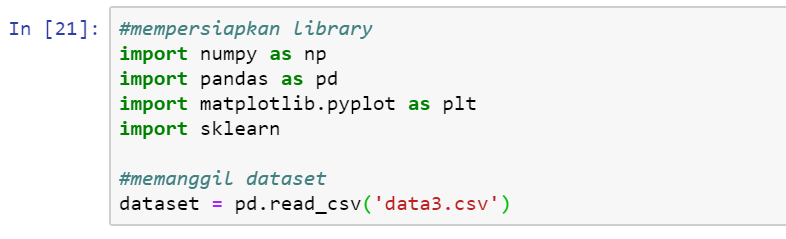
1. Membuat ploting data dengan memberi nama variabel sumbu x dan sumbu y, seperti dibawah ini



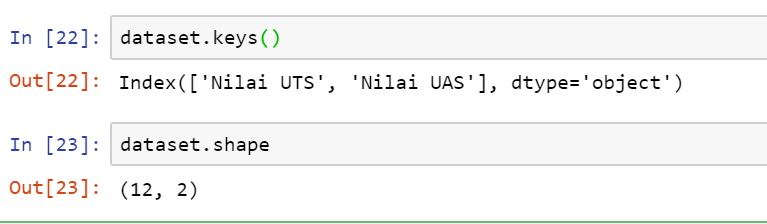
1. Manajemen data file csv
2. Menampilkan data 🡪 Buatlah data seperti dibawah ini dan simpan dengan nama data3.csv

|  |  |
| --- | --- |
| Nilai UTS | Nilai UAS |
| 80 | 87 |
| 78 | 81 |
| 72 | 70 |
| 68 | 65 |
| 78 | 90 |
| 74 | 78 |
| 90 | 87 |
| 76 | 78 |
| 73 | 81 |
| 66 | 88 |
| 89 | 94 |
| 75 | 86 |

Mendefiniskan library yang akan kita gunakan. Data disimpan dengan nama file data3.csv dalam satu folder dengan file python yang sedang running. Kita tampilkan menggunakan pandas dengan variabel dataset.



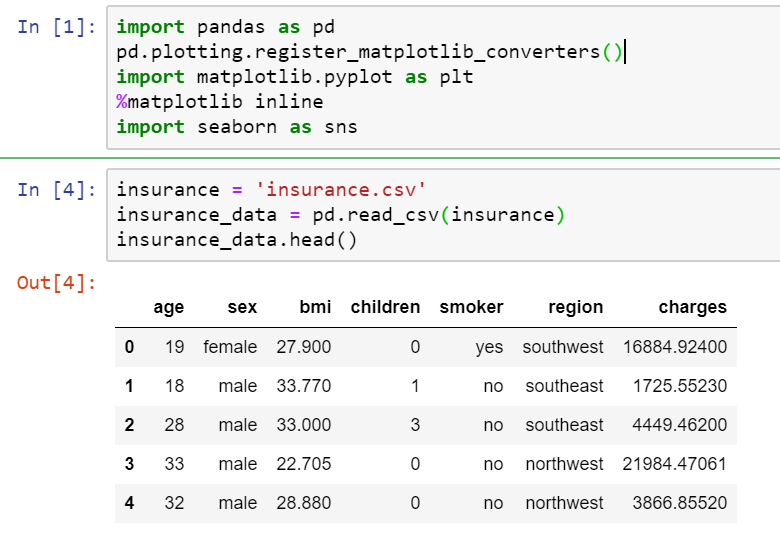
Untuk memastikan data tersimpan dan dapat dibaca oleh python gunakan script dibawah ini



Input [22] digunakan untuk menampilkan nama variabel pada data3.csv

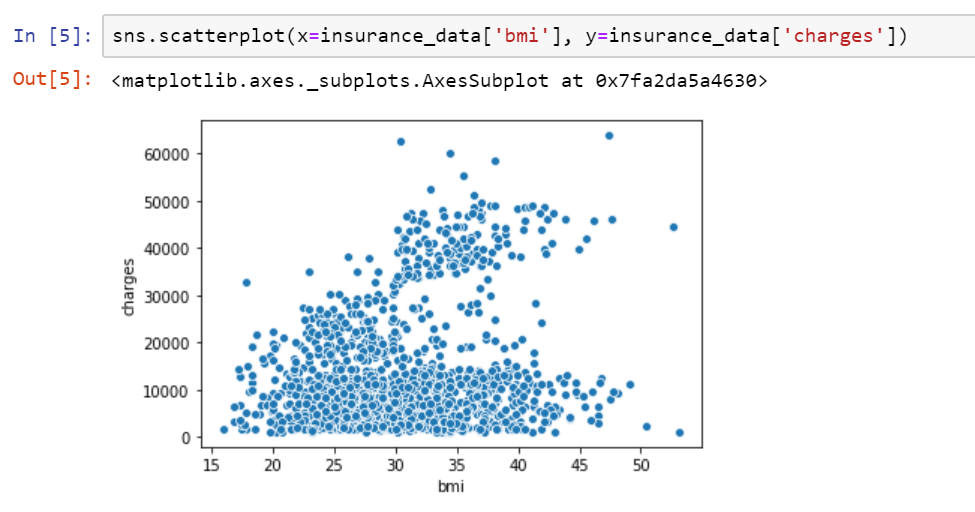
Input [23] digunakan untuk melihat jumlah record dan jumlah atribut dari data3.csv

1. Scatter Plot Data

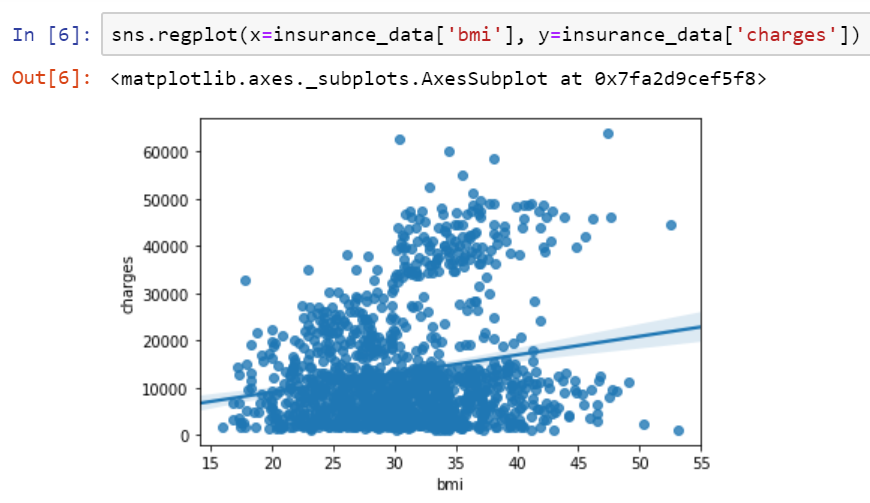


Apakah semua data bisa ditampilkan? Jelaskan!

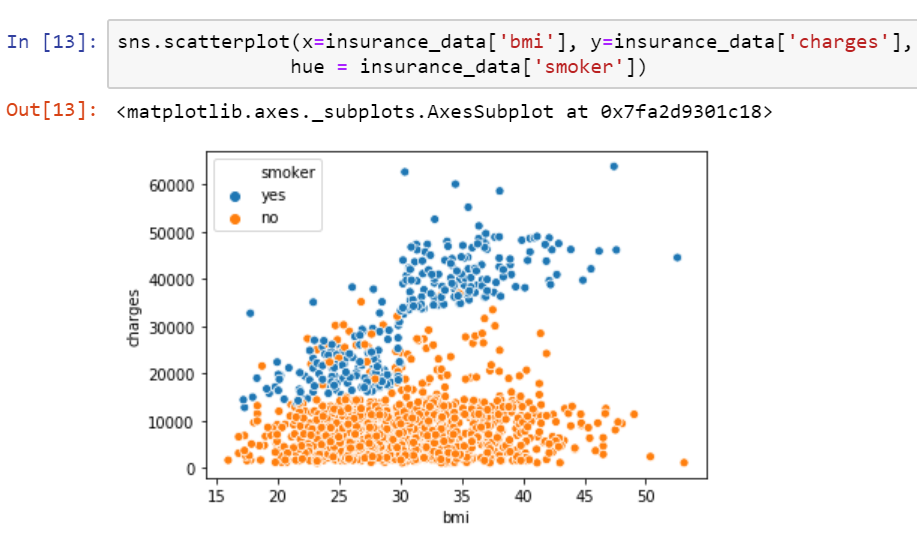
Fungsi untuk membuat scatter plot



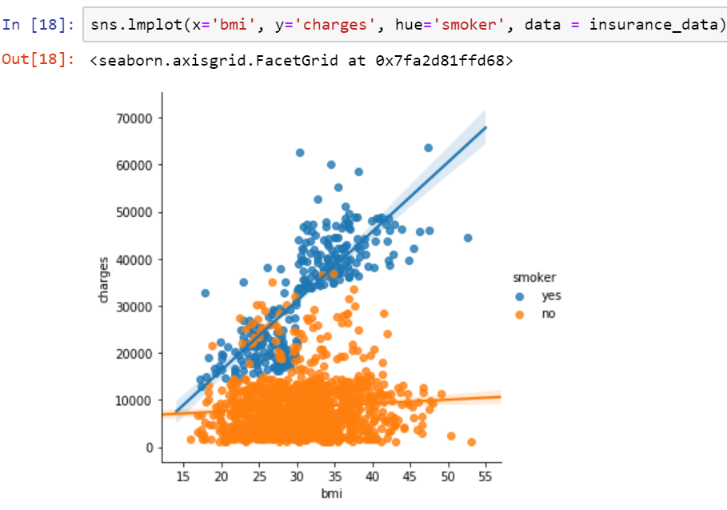
Untuk membuat plotting regresi dengan variabel x = bmi dan y = charges



Menampilkan plot data kategori dari salah satu variabel smoker



Menampilkan prediksi data tentang smoker melalui garis regresi



|  |  |
| --- | --- |
|  | LATIHAN |

1. Jelaskan perbedaan scatterplot, regplot, lmplot
2. Berikan contoh penggunaan fungsi yang dimiliki library Pandas untuk menampilkan data (minimal 8 fungsi)

|  |  |
| --- | --- |
| tugas.png | TUGAS |

1. Seaborn merupakan library yang dibangun diatas matplotlib dimana memiliki fungsi yang sama yaitu visualisasi data. Buatlah penjelasan tentang apa itu seaborn? Mengapa perlu menggunakan seaborn? Berikan contoh penggunaan seaborn selain yg ada di modul (minimal 2 fungsi).

|  |  |
| --- | --- |
|  | REFERENSI |