





MICROCREDENTIAL: ASSOCIATE DATA SCIENTIST

01 November – 10 Desember 2021

Pertemuan ke-4

Tools Proyek Data Science









Profil Pengajar: Nama Lengkap dan Gelar Akademik



Contak Pengajar:

Ponsel:

XXXXXX

Email:

XXXXXX

Jabatan Akademik:

Latar Belakang Pendidikan:

- S1:
- S2:
- S3:

Riwayat/Pengalaman Pekerjaan:

- Dosen
- Xxxx
- Xxxx
- Xxxx
- XXXX





Deskripsi Pelatihan

Tujuan utama dari modul pelatihan ini adalah untuk membahas data science tools dengan menjelaskan seperangkat kakas dan teknik yang berkaitan dengan keterampilan dasar dalam ilmu komputer, matematika, dan statistik untuk melakukan tugas-tugas yang umumnya terkait dengan data science.



Capaian Pembelajaran

Pada topik ini, kita akan mempelajari:

- Bahasa Pemrograman Python
- Development Environment
- Dasar-dasar library Python untuk proyek data science
 - NumPy
 - SciPy
 - **Pandas**
 - Matplotlib
 - Seaborn
 - Scikit-learn





Mengapa Python?



- Bahasa pemrograman tingkat tinggi
- Penulisan kode/sintaks lebih sederhana dan tersedia banyak library
- Bersifat open-source dan cross-platform
- Diluncurkan oleh Guido Van Rosum pada tahun 1991.

Data Professional



- Data Analyst
- Data Engineer
- Data Scientist
- Business Intelligence
- ML Engineer



- Cocok untuk pemula
- Sederhana tapi powerful
- High-demand skill





Mengapa Python?



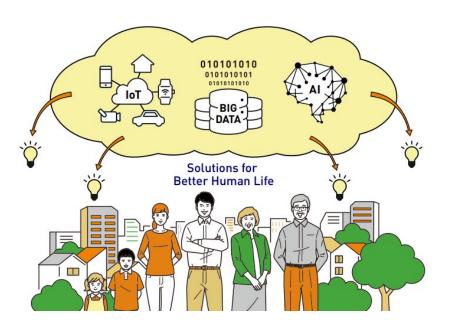
Keunggulan Python:

- Keterbacaan, Python mudah dibaca dan dipahami
- Efisien, memiliki library yang lengkap sehingga penulisan coding dapat lebih sederhana
- Multifungsi, dengan menggunakan Python dapat membuat website, aplikasi bidang robotika, aplikasi bidang kecerdasan artifisial.
- Interoperabilitas, Python mampu berinteraksi dengan Bahasa pemrograman lainnya
- Dukungan komunitas, Python merupakan program open source, dan komunitas Python sangat aktif dalam melakukan pengembangan

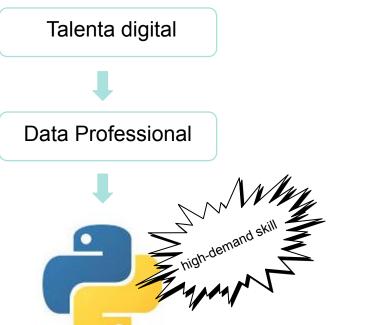




Mengapa Python?



https://www.japan.go.jp/abenomics/_userdata/abenomics/pdf/society_5.0.pdf

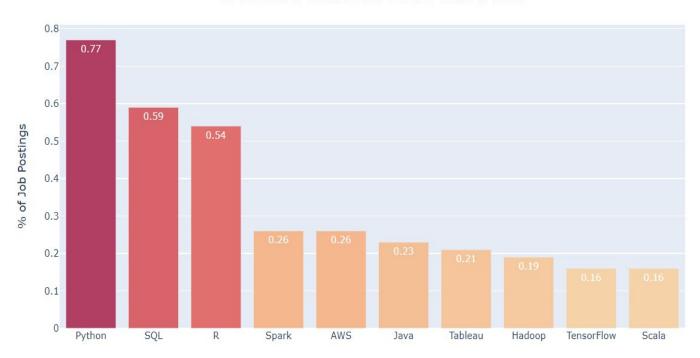






Python Menjadi yang Pertama dalam Daftar Keahlian yang Paling Dibutuhkan

10 Most In-Demand Data Science Skills in 2021



(sumber: https://towardsdatascience.com)







Python Digunakan pada YouTube

"Google runs millions of lines of Python code. The front-end server that drives youtube.com and YouTubes APIs is primarily written in Python, and it serves millions of requests per second!"

Dylan Trotter, Youtube Engineer, 2017

You Tube

https://opensource.googleblog.com/2017/01/grumpy-go-running-python.html





Python Digunakan pada Quora



"We decided that Python was fast enough for most of what we need to do (since we push our performance-critical code to backend servers written in C++ whenever possible). As far as type checking, we ended up writing very thorough unit tests which are worth writing anyway, and achieve most of the same goals."

- Adam D'Angelo, CEO Quora, 2014

https://www.quora.com/Why-did-Quora-choose-Python-for-its-development





Python Digunakan pada Beberapa Industri







































Penerapan Python pada Proyek Data Science



Data Exploration

Data Pre-Processing

Data Cleansing

Data Modeling

- Scraping, crawling, data mining
- Coding, query
- Seleksi fitur, statistika deskriptif, class balancing, visualisasi data
- Transformasi fitur: Categorical encoding, binning
- Menangani nilai kosong (missing values), menghapus baris terduplikasi
- Data formatting, menangani data pencilan (outliers)
- Melatih data dengan algoritma machine learning
- Melakukan klasifikasi, regresi, prediksi, klasterisasi





Memulai Python

- **Python** adalah bahasa *interpreter*, yang dapat mengurangi siklus *edit-test-debug* karena tidak memerlukan langkah kompilasi
- Untuk menjalankan Python, Anda memerlukan *runtime/interpreter environment* untuk mengeksekusi kode:
 - Mode interaktif: Setiap perintah yang Anda tulis akan langsung ditafsirkan dan segera dieksekusi sehingga bisa langsung melihat hasilnya □ **IPython**
 - Mode skrip: Anda memasukkan satu set kode Python ke dalam format .py, program dijalankan baris demi baris



Konsep IPython: REPL Environment

Read

Proses membaca code

Eval

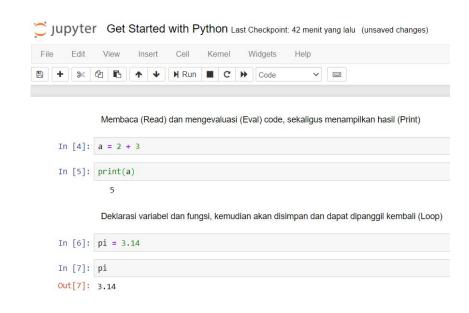
 Proses evaluasi (eksekusi) code



Proses menampilkan hasil (output)



Pengulangan proses R-E-P



Pilihan Development Environment

Pilih *Development Environment* yang paling mudah dan nyaman:

- Anaconda Distribution (<u>https://www.anaconda.com/distribution/</u>)
 - Python
 - Conda lebih dari 1000 library data science
- Miniconda (https://docs.conda.io/en/latest/miniconda.html)
 - Python interpreter
 - Conda
- Jupyter Notebook (<u>https://jupyter.org/</u>)
- Python installer (<u>https://www.python.org/downloads/</u>).
- Google Colaboratory (<u>https://colab.research.google.com/</u>).
- Notebooks Azure (<u>https://notebooks.azure.com/</u>)
- Github (https://github.com/)





Anaconda Distribution

Jupyter Notebook



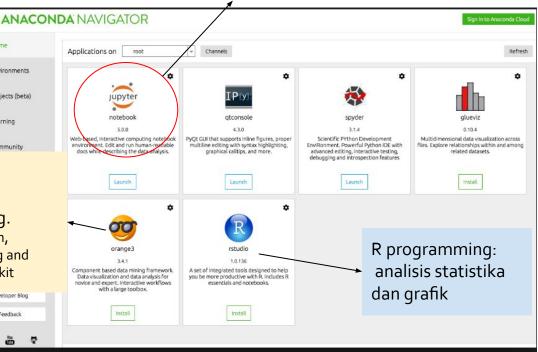


Anaconda Navigator

Sebuah aplikasi dashboard interface pada paket Anaconda Distribution



Visual





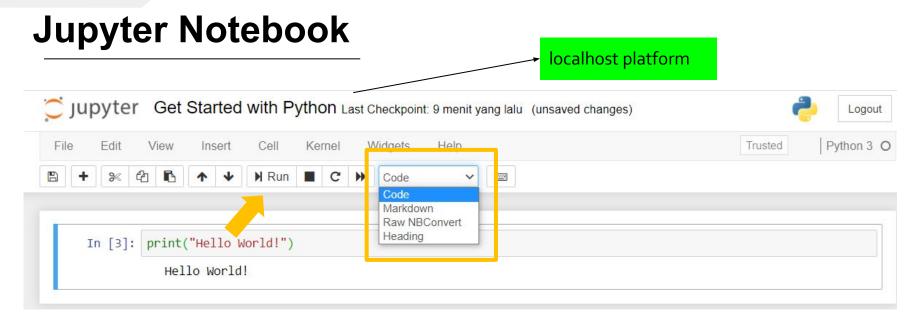


Jupyter Notebook

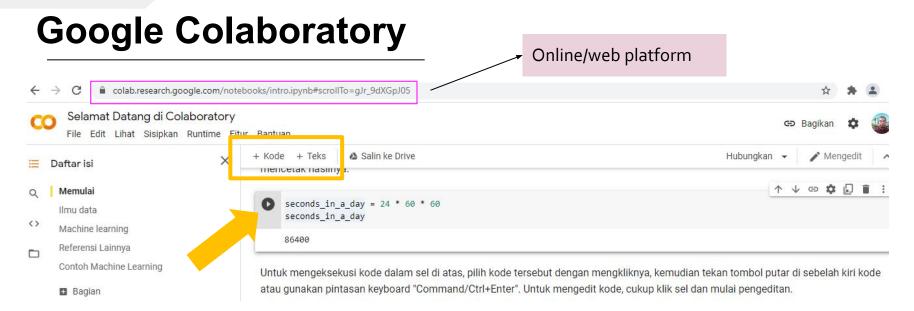


- Lingkungan pemrograman interaktif berbasis web yang mendukung berbagai bahasa pemrograman termasuk Python
- Banyak digunakan oleh peneliti dan akademisi untuk pemodelan matematika, pembelajaran mesin, analisis statistik, dan untuk pengajaran pemrograman





- Skrip dapat ditulis dalam bentuk:
 - Code : Algoritma dan formula matematis
 - o *Markdown/Heading*: Teks deskripsi, penjelasan code, gambar/image
 - o Raw NBConvert: Konversi format yang berbeda
- Hasil dapat diketahui langsung setelah menjalankan perintah Run



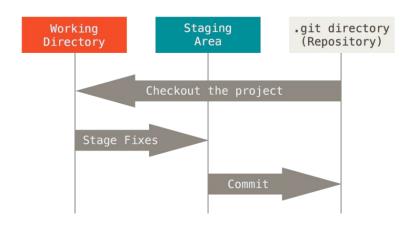
- Skrip dapat ditulis dalam bentuk:
 - Code: Algoritma dan formula matematis
 - Teks: Teks deskripsi, penjelasan code
- Dapat digunakan pada https://colab.research.google.com/ dan hasil dapat diketahui langsung setelah menjalankan perintah Run





Bekerja dengan Git

- Git merupakan tools yang bersifat open source untuk memudahkan bekerja dengan proyek berskala kecil maupun besar (https://git-scm.com/)
- Git memiliki tiga status utama tempat file berada: modified, staged, committed:
 - Modified berarti Anda telah mengubah file tetapi belum menyimpannya ke database Anda
 - Staged berarti Anda telah menandai file yang dimodifikasi dalam versi terbaru untuk masuk ke tahap commit
 - Commit berarti bahwa data disimpan dengan aman di database local Anda







Bekerja dengan Git

- Inisialisasi: git init
- **Commit**: git commit -m "first commit"
- **Branch**: git branch -M main
- Add: git remote add origin https://github.com/[user]/[repo].git
- **Push**: git push -u origin main
- **Pull**: git pull origin [branch]





Hello World!

Bahasa C

```
#include <stdio.h>
int main() {
printf("Hello
World!");
return 0;
}
```

Bahasa Python

```
print("Hello
World!")
```

- Lebih sederhana
- Tidak ada kurung kurawal {..}
- Tidak perlu titik koma;





Tipe Data Python

- float bilangan riil
- int bilangan bulat (integer)
- str string, teks
- bool True or False

```
In [1]: height = 1.84
In [2]: tall = True
```

Masalah

- Terlalu banyak data masukan untuk tipe data yang sama
- Tidak nyaman

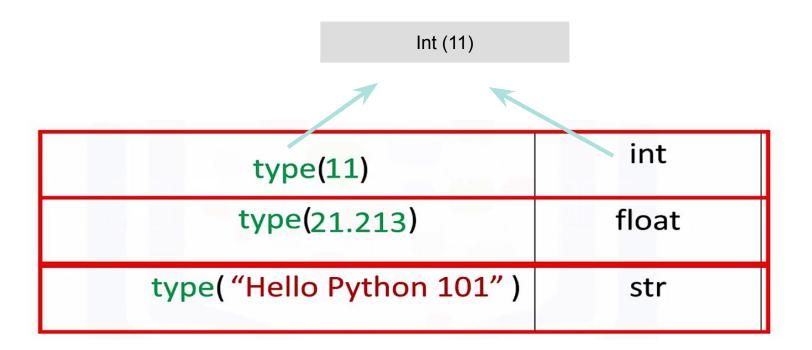
```
In [3]: height1 = 1.84
In [4]: height2 = 1.79
In [5]: height3 = 1.82
In [6]: height4 = 1.90
```

Solusi □ Python List





Т



Tipe Data Python

T

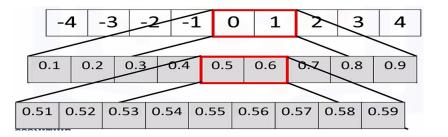
Tipe Data Integer

Merupakan bilangan bulat negatif maupun positif



Tipe Data Float

Bilangan bentuk bilangan pecahan dapat negatif maupun positif



Tipe Data String

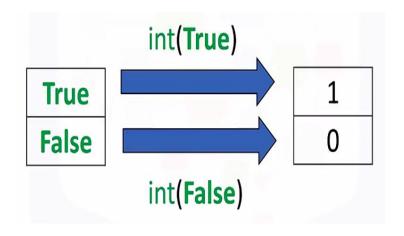
Merupakan data yang berisi string berupa angka, huruf dan symbol.

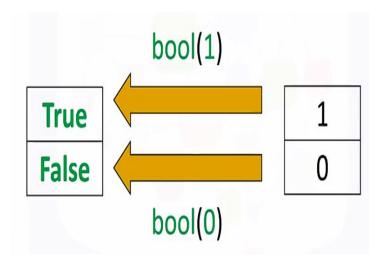


Tipe Data Boolean

Merupakan tipe data dengan dua keluaran yaitu: True atau False.

Pada Python, tipe data boolean adalah **bool**.





Type Casting

Pada Python, dapat dilakukan pengubahan tipe data.

Misalkan: tipe data integer akan diubah menjadi tipe data float.



List

List merupakan daftar yang memiliki anggota/elemen yang berurutan. Anggota/elemen dari list dituliskan di antara tanda kurung siku [...] dan masing-masing anggota/elemen dipisahkan oleh tanda koma (,). List bersifat **mutable** (tidak tetap).

List dapat berisikan anggota/elemen dengan tipe data string, integer, float, bersarang.



List

Anggota/elemen pada list dapat diakses dengan menggunakan indeks. Indeks pada list dapat dibuat dalam bentuk indeks positif maupun indeks negatif.

L	=["Michael Jackson", 10.1,198				
	0	"Michael Jackson"	L[0]: "Michael Jackson"		
	1	10.1	L[1]: 10		
	2	1982	L[2]: 1982		

-3	0	"Michael Jackson"	L[-3]: "Michael Jackson"
-2	1	10.1	L[-2]: 10
-1	2	1982	L[-1]: 1982

L =["Michael Jackson", 10.1,1982]

Indeks Positif

Indeks Negatif

Python List [a, b, c]

- Koleksi nilai
- Dapat mengandung beberapa tipe data berbeda

```
In [7]: [1.84, 1.79, 1.82, 1.90, 1.80]
Out[7]: [1.84, 1.79, 1.82, 1.90, 1.80]
In [8]: height = [1.84, 1.79, 1.82,
1.90, 1.80]
In [9]: height
Out[9]: [1.84, 1.79, 1.82, 1.90, 1.80]
```

```
In [10]: famz = ["Abe", 1.84, "Beb",
1.79, "Cory", 1.82, "Dad", 1.90]
In [11]: famz
Out[11]: ["Abe", 1.84, "Beb", 1.79,
 "Cory", 1.82, "Dad", 1.90]
              ["Abe", 1.84]
              ["Beb", 1.79]
             ["Cory", 1.82]
              ["Dad", 1.90]
```



Python List

```
In [1]: height = [1.84, 1.79, 1.82, 1.90, 1.80]
In [2]: height
Out[2]: [1.84, 1.79, 1.82, 1.90, 1.80]
In [3]: weight = [66.5, 60.3, 64.7, 89.5, 69.8]
In [4]: weight
Out[4]: [66.5, 60.3, 64.7, 89.5, 69.8]
                                                         Problem!
In [5]: weight / height ** 2
TypeError: unsupported operand type(s) for ** or pow(): 'list' and
'int.'
```

Solusi:

NumPy **

- Library dasar untuk perhitungan saintifik (scientific computing) dengan Python (https://numpy.org/)
- Alternatif untuk Python List: Numpy Array untuk n-dimensi
- Mudah digunakan dan bersifat open source
- Jika library belum terpasang, tuliskan perintah instalasi:

pip install numpy

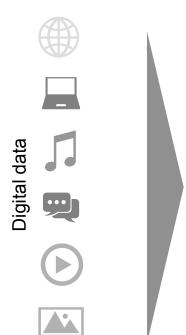
Kemudian impor:

```
import numpy as np
```

```
In [6]: import numpy as np
In [7]: np height = np.array(height)
In [8]: np height
Out[8]: array([1.84, 1.79, 1.82, 1.9, 1.8])
In [9]: np weight = np.array(weight)
In [10]: np weight
Out[10]: array([66.5, 60.3, 64.7, 89.5, 69.8])
In [11]: bmi = np weight / np height ** 2
In [12]: bmi
Out[12]: array([19.64201323, 18.81963734,
19.53266514, 24.79224377, 21.54320988])
```







- Pengolahan data dapat berupa bermacam-macam bentuk dan formatnya: dokumen, gambar, video, suara, angka, atau teks
- Ketika data-data tersebut diproses, tidak secara mentah-mentah dibaca sebagai video atau audio. Tetapi sudah dilakukan transformasi ke dalam bentuk array atau matrix of number
- Array dengan minimal dua dimensi akan membentuk matriks dan dapat menggunakan NumPy

```
import numpy as np
np.<TAB>
```







 NumPy juga dapat digunakan untuk membuat array berdimensi-n

```
In [13]: import numpy as np
In [14]: np_height = np.array([1.84, 1.79,
1.82, 1.9, 1.8])
In [15]: np_weight = np.array([66.5, 60.3,
64.7, 89.5, 69.8])
In [16]: type(np_height)
Out[16]: numpy.ndarray
In [16]: type(np_weight)
Out[16]: numpy.ndarray
```

```
ndarray = n-dimensional array
```

Array berdimensi 2 baris 5 kolom \rightarrow Matriks M_{2x5}

$$M = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \end{bmatrix}$$







- **SciPy** (dibaca "Sigh Pie") merupakan library yang bersifat open source dan tersedia di https://www.scipy.org/
- SciPy dibangun untuk untuk bekerja dengan NumPy array dan menyediakan kumpulan algoritma numerik, termasuk pemrosesan sinyal, optimasi, statistika, dan library Matplotlib untuk visualisasi data.
- Jika library belum terpasang, tuliskan perintah instalasi:

pip install scipy







- Pandas (Panel Data) merupakan library popular di Python yang digunakan untuk data structure dan data analysis
- Bersifat open source dan tersedia di https://pandas.pydata.org/
- Pandas sangat berkaitan dengan NumPy
- Jika library belum terpasang, tuliskan perintah instalasi:

pip install pandas

Kemudian impor:

import pandas as pd

Data Wrangling / Data Munging

- Reshaping (mengubah bentuk data)
- Joining (menggabungkan data)
- Splitting (pemisahan data)
- Time-series analysis (data berkala)

Data Cleansing

- Membersihkan data tidak lengkap (Error)
- Menangani data pencilan (outliers)
- Menghapus data duplikat







Representasi Data di



- Terdapat 2 data objects: Series dan DataFrame
- Series □ Data berbentuk 1 dimensi

```
In [13]: np.array([1, 2, 3, 4, 5])
Out[13]: array([1, 2, 3, 4, 5])
```

 DataFrame □ Data berbentuk 2 dimensi atau lebih

Kolom: Fitur / atribut/variabel

Baris: sampel

	Negara	Populasi	Area	Ibukota
IN	Indonesia	250	123456	Jakarta
MA	Malaysia	25	3456	KL
SI	Singapura	15	456	Singapura
JP	Jepang	60	5678	Tokyo
TH	Thailand	45	678	Bangkok







- Pandas dapat mengimpor data dari berbagai format: comma-separated value (CSV), file teks, Microsoft Excel, database SQL, dan format HDF5
- Unduh dataset: http://bit.ly/TabDataset
- CSV file □ DataFrame import pandas as pd

Tab.csv

Negara,Populasi,Area,Ibukota IN,Indonesia,250,123456,Jakarta MA,Malaysia,25,3456,KL SI,Singapura,15,456,Singapura JP,Jepang,60,5678,Tokyo TH,Thailand,45,678,Bangkok

In [1]: Tab =	# deklarasi tabel
In [2]: Tab	

	Negara	Populasi	Area	Ibukota
IN	Indonesia	250	123456	Jakarta
MA	Malaysia	25	3456	KL
SI	Singapura	15	456	Singapura
JP	Jepang	60	5678	Tokyo
TH	Thailand	45	678	Bangkok







In [3]: import pandas as pd
In [4]: Tab = pd.read_csv("Tab.csv")
In [5]: Tab
Out[5]:

	Unnamed: 0	Negara	Populasi	Area	lbukota
0	IN	Indonesia	250	123456	Jakarta
1	MA	Malaysia	25	3456	KL
2	SI	Singapura	15	456	Singapura
3	JP	Jepang	60	5678	Tokyo
4	TH	Thailand	45	678	Bangkok

In [6]: Out[6]:	Tab["Negara"]	# akses kolom
	<pre>0 Indonesia 1 Malaysia 2 Singapura 3 Jepang 4 Thailand Name: Negara, dtype:</pre>	object
In [7]: Out[7]:	Tab.Ibukota	# akses kolom
	<pre>0 Jakarta 1 KL 2 Singapura 3 Tokyo 4 Bangkok Name: Ibukota, dtype:</pre>	: object



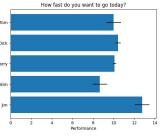


- Matplotlib adalah library Python untuk visualisasi data dengan dua dimensi
- Bersifat open source dan tersedia di https://matplotlib.org/
- Matplotlib berkaitan dengan NumPy dan Pandas
- Jika library belum terpasang, tuliskan perintah instalasi:

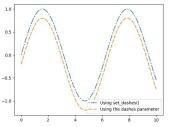
pip install matplotlib

Kemudian impor:

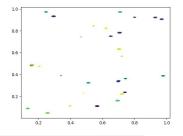
import matplotlib.pyplot as plt



bar chart



Line chart



Scatter plot

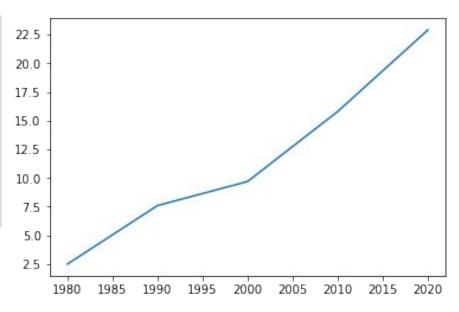








```
In [1]: import matplotlib.pyplot as plt
In [2]: year = [1980, 1990, 2000, 2010, 2020]
In [3]: price = [2.5, 7.6, 9.7, 15.8, 22.9]
In [4]: plt.plot(year, price)
In [5]: plt.show()
```

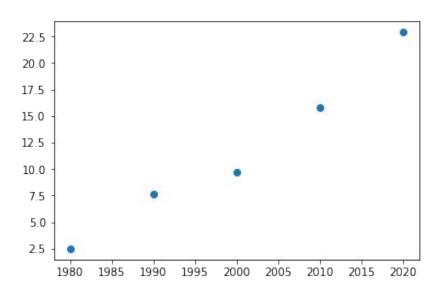




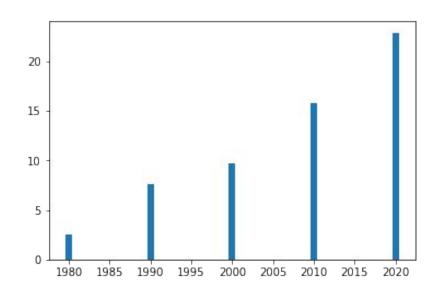




In [6]: plt.scatter(year, price)



In [7]: plt.bar(year, price)









- Seaborn adalah library visualisasi data Python (serupa dengan Matplotlib) yang menyediakan high-level interface untuk menggambar grafik statistika yang menarik dan informatif
- Library ini bersifat open source dan tersedia di https://seaborn.pydata.org/
- Jika library belum terpasang, tuliskan perintah instalasi:

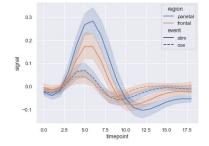
pip install seaborn

Kemudian impor:

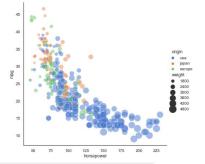
import seaborn as sns







Line chart



Scatter plot







- Scikit-learn adalah library untuk mempraktikkan machine learning dan membuat model
- Bersifat open source dan tersedia di https://scikit-learn.org/
- Scikit-learn diawali dari project SciPy (Scientific Python) yang berisi fungsi-fungsi matematis
- Jika library belum terpasang, tuliskan perintah instalasi:

pip install sklearn

Kemudian impor:

import sklearn

Classification

- Support Vector Machines
- Decision Tree
- Random Forest
- Neural Network
- Nearest neighbors

Clustering

- K-Means Clustering
- Hierarchical Clustering

Model Selection

- Cross validation
- Metrics







Summary

Pada topik ini, kita sudah mempelajari:

- Keunggulan Python sebagai Tools dalam proyek data science
- Development environment Python yang bervariasi, baik yang bersifat offline (local computer) maupun berbasis web (Jupyter Notebook / Google Colaboratory)
- Dasar-dasar library Python untuk proyek data science:
 - NumPy

 library untuk numerical computation
 - SciPy □ library untuk perhitungan statistika dan matematis
 - Pandas

 library untuk analisis dan manipulasi data
 - Matplotlib ☐ library untuk visualisasi data
 - Seaborn □ library untuk visualisasi data dengan *high-level interface*
 - Scikit-learn

 library untuk mempraktikkan *machine learning* dan pemodelan

Tools / Lab Online

- Python for Data Science: https://cognitiveclass.ai/courses/python-for-data-science
- Introduction to Python for Data Science: https://www.datacamp.com/courses/intro-to-python-for-data-science
- Build real-world applications with Python: https://docs.microsoft.com/en-us/learn/paths/python-language/
- Introduction to Python for Data Science: https://learning.edx.org/course/course-v1:Microsoft+DAT208x+3T2018/home
- Data Visualization with Python: https://cognitiveclass.ai/courses/data-visualization-with-python



Referensi

- Arfika Nurhudatiana. 2019. Analisis dan Visualisasi Data Dengan Python. Pintaria.
- Introduction to Python for Data Science:
 https://learning.edx.org/course/course-v1:Microsoft+DAT208x+3T2018/home
- Jake VanderPlas. 2016. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
- Laura Igual and Santi Segui. 2017. *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Springer International Publishing.
- Python Tutorial: Learn Python for Data Science (Youtube: DataCamp).
 https://www.youtube.com/watch?v=-Rf4fZDQ0yw&list=PLjgj6kdf_snaw8QnlhK5f3DzFDFKDU5f4



Quiz / Tugas

Quiz dapat diakses melalui https://spadadikti.id/



48

Terima kasih



