Hands-On

Hands-On ini digunakan pada kegiatan Microcredential Associate Data Scientist 2021

Pertemuan 5

Pertemuan 5 (lima) pada Microcredential Associate Data Scientist 2021 menyampaikan materi mengenai Mengumpulkan Data, Menelaah Data dengan metode Statistik

Pengambilan Data dari API Kaggle

Salah satu portal yang menyediakan dataset untuk project Data Science adalah Kaggle (https://www.kaggle.com/). Pada latihan ini, silakan peserta mengunduh dataset mengenai bunga Iris dengan menggunakan kata kunci: "iris species" yang disediakan oleh UCI Machine Learning (UCIML)

1. Install Modul kaggle:

In [91]:

```
# Install modul kaggle secara inline (di dalam notebook)
!pip install kaggle
```

```
Requirement already satisfied: kaggle in c:\users\shinyq\appdata\local\programs\python\python39\lib\site-packag
es (1.5.12)
Requirement already satisfied: six>=1.10 in c:\users\shinyq\appdata\local\programs\python\python39\lib\site-pac
kages (from kaggle) (1.16.0)
Requirement already satisfied: certifi in c:\users\shinyq\appdata\local\programs\python\python39\lib\site-packa
ges (from kaggle) (2020.11.8)
Requirement already satisfied: python-dateutil in c:\users\shinyq\appdata\local\programs\python\python39\lib\si
te-packages (from kaggle) (2.8.2)
Requirement already satisfied: requests in c:\users\shinyq\appdata\local\programs\python\python39\lib\site-pack
ages (from kaggle) (2.21.0)
Requirement already satisfied: tqdm in c:\users\shinyq\appdata\local\programs\python\python39\lib\site-packages
(from kaggle) (4.62.1)
Requirement already satisfied: python-slugify in c:\users\shinyq\appdata\local\programs\python\python39\lib\sit
e-packages (from kaggle) (4.0.1)
Requirement already satisfied: urllib3 in c:\users\shinyq\appdata\local\programs\python\python39\lib\site-packa
ges (from kaggle) (1.24.3)
Requirement already satisfied: text-unidecode>=1.3 in c:\users\shinyq\appdata\local\programs\python\python39\li
b\site-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: idna<2.9,>=2.5 in c:\users\shinyq\appdata\local\programs\python\python39\lib\sit
e-packages (from requests->kaggle) (2.8)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in c:\users\shinyq\appdata\local\programs\python\python39
\lib\site-packages (from requests->kaggle) (3.0.4)
Requirement already satisfied: colorama in c:\users\shinyq\appdata\local\programs\python\python39\lib\site-pack
ages (from tqdm->kaggle) (0.4.4)
WARNING: You are using pip version 21.2.3; however, version 21.3.1 is available.
You should consider upgrading via the 'c:\users\shinyq\appdata\local\programs\python\python39\python.exe -m pip
install --upgrade pip' command.
```

In [92]:

```
# Install modul kaggle secara eksternal melalui anaconda prompt:
```

2. Create Token API kaggle:

1. Login Kaggle.com
2. Kemudian pada menu Profile --> Account
3. Klik Create New Api Token
4. Maka akan terdownload file kaggle.json

Kaggle API secara default mengasumsikan bahwa file kaggle.json tersebut berada di dalam folder:

- ~/.kaggle/ (Linux/Mac)
- C:\Users\.kaggle\ (Windows)

Jika folder tersebut belum ada:

1. Buat folder di direktori C:\Users\.kaggle\
2. letakkan file kaggle.json kedalam folder tersebut

3. Download Dataset dari Kaggle:

Dokumentasi Kaggle Commands selengkapnya Disini

In [93]:

```
# Mencari dataset yang tersedia di kaggle --> pilih data provider dari UCIML
!kaggle datasets list -s Iris
```

```
ref                                                   title                                             size  l
astUpdated          downloadCount  voteCount  usabilityRating
----------------------------------------------------  ---------------------------------------------  -----  -
-------------------  -------------  ---------  ---------------
uciml/iris                                            Iris Species                                       4KB  2
016-09-27 07:38:05         226819       2680  0.7941176
arshid/iris-flower-dataset                            Iris Flower Dataset                              1010B  2
018-03-22 15:18:06          40650        371  0.8235294
vikrishnan/iris-dataset                               Iris Dataset                                      999B  2
017-08-03 16:00:44           2934         26  0.7647059
therohk/ireland-historical-news                       Irish Times - Waxy-Wany News                      52MB  2
021-09-25 10:52:48           2984        157  1.0
chuckyin/iris-datasets                                Iris datasets                                      1KB  2
017-03-10 09:35:43           1774         14  0.7352941
rtatman/iris-dataset-json-version                     Iris Dataset (JSON Version)                        1KB  2
018-04-06 20:21:31           5639         43  0.75
parulpandey/palmer-archipelago-antarctica-penguin-data  Palmer Archipelago (Antarctica) penguin data    11KB  2
020-06-09 10:14:54          10083        115  0.9705882
conorrot/irish-weather-hourly-data                    Irish Weather (hourly data)                       67MB  2
020-06-29 20:15:18           1867         40  0.8235294
saurabh00007/iriscsv                                  Iris.csv                                           1KB  2
017-11-09 07:34:35          17167         57  0.4117647
jillanisofttech/iris-dataset-uci                      Iris dataset uci                                   1KB  2
021-11-06 15:11:47             37         12  1.0
fleanend/birds-songs-numeric-dataset                  Birds' Songs Numeric Dataset                      25MB  2
019-04-01 09:09:46            706         25  0.9411765
kamrankausar/iris-data                                iris_data                                          1KB  2
017-11-30 10:26:01           1120         13  0.64705884
jeffheaton/iris-computer-vision                       Iris Computer Vision                               5MB  2
020-11-24 21:23:29            312          9  0.875
styven/iris-dataset                                   Iris dataset                                       1KB  2
017-11-04 14:10:12            797          8  0.29411766
arslanali4343/iris-species                            Iris Species                                       2KB  2
020-07-02 06:09:09             61         13  0.5625
olgabelitskaya/flower-color-images                    Flower Color Images                               50MB  2
020-10-01 22:48:07           8375        161  0.75
naureenmohammad/mmu-iris-dataset                      MMU iris dataset                                  30MB  2
020-07-25 18:38:33            645         19  0.5625
rutujavaidya/iris-dataset                             Iris Dataset                                       1KB  2
021-07-25 17:37:14             36          6  0.4117647
shantanuss/iris-flower-dataset                        IRIS flower dataset                                1KB  2
020-01-18 19:43:18            200          3  0.9411765
ashishs0ni/iris-dataset                               Iris_dataset                                       1KB  2
018-08-05 14:26:19            602          7  0.64705884
```

In [94]:

```
# Download dan ekstrak dataset, secara default akan berada dalam satu direktori dengan notebook ini
!kaggle datasets download uciml/iris --unzip
```

```
Downloading iris.zip to c:\Users\ShinyQ\Desktop\Tugas Mandiri Microcredential\Pertemuan 5

  0%|          | 0.00/3.60k [00:00<?, ?B/s]
100%|██████████| 3.60k/3.60k [00:00<00:00, 3.69MB/s]
```

Atau bisa juga menggunakan link dari kaggle

Latihan (1)

Silahkan Download sebuah dataset menggunakan API Kaggle

In [95]:

```
#Latihan (1)
#Langkah nya seperti contoh diatas
!kaggle datasets download jeffheaton/iris-computer-vision --unzip
```

```
Downloading iris-computer-vision.zip to c:\Users\ShinyQ\Desktop\Tugas Mandiri Microcredential\Pertemuan 5

  0%|          | 0.00/5.33M [00:00<?, ?B/s]
 19%|██        | 1.00M/5.33M [00:00<00:00, 5.50MB/s]
 38%|████      | 2.00M/5.33M [00:00<00:00, 5.05MB/s]
 56%|██████    | 3.00M/5.33M [00:00<00:00, 4.83MB/s]
 75%|████████  | 4.00M/5.33M [00:00<00:00, 4.92MB/s]
 94%|██████████ | 5.00M/5.33M [00:01<00:00, 5.10MB/s]
100%|██████████| 5.33M/5.33M [00:01<00:00, 5.16MB/s]
```

PENGGUNAAN LIBRARY PANDAS dan NUMPY

Pada materi ini, peserta sudah mendapatkan pemahaman mengenai data dan dataset. Penggunaan library pada Python memberikan kemudahan dalam proses data understanding. Beberapa library yang digunakan adalah library Pandas dan Numpy.

Latihan (2)

Lakukan import Library Pandas dan Library Numpy

In [96]:

```
#Latihan(2)
#Import Library Pandas
import pandas as pd

#Import Library Numpy
import numpy as np
```

DATAFRAME

DataFrame adalah struktur data 2 dimensi yang berbentuk tabular (mempunyai baris dan kolom). Hampir semua data tidak hanya memiliki 1 kolom tetapi lebih dari 1 kolom, sehingga lebih cocok menggunakan pandas DataFrame untuk mengolahnya.

Penggunaan dataframe pada Python dengan menggunakan syntaks: df.

Latihan (3)

Panggil file (load dataset) dengan format .csv untuk dataset mengenai bunga Iris yang sudah peserta unduh dari Kaggle, dan akan disimpan di dalam dataframe df. Lalu tampilkan 5 baris awal dataset dengan function head()

In [97]:

```
#latihan(3)
#Panggil file (load file bernama Iris.csv) dan simpan dalam dataframe Lalu tampilkan 5 baris awal dataset denga
df = pd.read_csv('Iris.csv')
df.head(5)
```

Out[97]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

Telaah Data

Pada telaah data, dapat dilakukan untuk mengetahui:

- tipe data dari setiap kolom
- deskripsi statistik data

Latihan (4)

Tampilkan tipe data dari kolom yang ada pada dataset

In [98]:

```
#latihan(4)
#Tampilkan tipe data dari kolom yang ada pada dataset
df.dtypes
```

Out[98]:

```
Id                 int64
SepalLengthCm    float64
SepalWidthCm     float64
PetalLengthCm    float64
PetalWidthCm     float64
Species           object
dtype: object
```

Latihan (5)

Apakah tipe Data dari kolom berikut ini: (silakan diisi pada cell di bawah ini)

In [99]:

```
#Latihan (5)
#Tipe Data dari kolom yang ada di dataset

# Kolom  "Id" memiliki tipe data = int64
# Kolom "SepalLengthCm" memiliki tipe data = float64
# Kolom "Species" memiliki tipe data = object
```

Latihan (6)

Hitunglah ukuran (jumlah baris dan kolom) dari dataset. Dengan menggunakan method function

In [100…

```
#Latihan (6)
#Hitung ukuran (jumlah baris dan kolom) dari dataset
df.shape
```

Out[100…

```
(150, 6)
```

Latihan (7)

Berapakah jumlah baris, dan jumlah kolom pada dataset? (silakan diisi pada cell di bawah ini)

In [101…

```
#Latihan (7)

#Jumlah Baris pada dataset adalah =  150

#Jumlah kolom pada dataset adalah =  6
```

Latihan (8)

Tampilkan data yang hanya berisi kolom "Id" dan kolom "Species" dalam bentuk dataframe.

In [102…

```
#Latihan (8)
#Tampilkan data untuk kolom "Id" dan kolom "Species" dalam bentuk dataframe
df[['Id', 'Species']]
```

Out[102…

	Id	Species
0	1	Iris-setosa
1	2	Iris-setosa
2	3	Iris-setosa
3	4	Iris-setosa
4	5	Iris-setosa
...	...	...
145	146	Iris-virginica
146	147	Iris-virginica
147	148	Iris-virginica
148	149	Iris-virginica
149	150	Iris-virginica

150 rows × 2 columns

Latihan (9)

Tampilkan data dengan dataframe, dan data yang ditampilkan adalah data pada baris dengan indeks 0 (nol) sampai dengan indeks 9 (sembilan)

In [103…

```
#Latihan (9)
#Tampilkan data dengan dataframe, dan data yang ditampilkan adalah baris dengan indeks 0 (nol) sampai dengan in

df.head(10)
```

Out[103…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
5	6	5.4	3.9	1.7	0.4	Iris-setosa
6	7	4.6	3.4	1.4	0.3	Iris-setosa
7	8	5.0	3.4	1.5	0.2	Iris-setosa
8	9	4.4	2.9	1.4	0.2	Iris-setosa
9	10	4.9	3.1	1.5	0.1	Iris-setosa

Latihan (10)

Tampilkan data hanya kolom "Id" dan kolom "Species" dengan dataframe, dan yang ditampilkan adalah data pada baris dengan indeks 11 (sebelas) sampai dengan indeks 15 (limabelas)

In [104…

```
#Latihan (10)
#Tampilkan data hanya kolom "Id" dan kolom "Species", pada baris dengan indeks 0 (nol) sampai dengan indeks 9

df[['Id', 'Species']].head(10)
```

Out[104…

	Id	Species
0	1	Iris-setosa
1	2	Iris-setosa
2	3	Iris-setosa
3	4	Iris-setosa
4	5	Iris-setosa
5	6	Iris-setosa
6	7	Iris-setosa
7	8	Iris-setosa
8	9	Iris-setosa
9	10	Iris-setosa

Latihan (11)

Pada DataFrame dapat menampilkan beberapa baris pertama/terakhir dari dataset yang di load. Gunakan Method head() dan tail().

Latihan: Tampilkan data pada 8 (delapan) baris pertama dari dataset, dengan dataframe.

In [105…

```
#Latihan (11)
#Tampilkan data pada 8 (delapan) baris pertama dari dataset, dengan dataframe

df.head(8)
```

Out[105…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
5	6	5.4	3.9	1.7	0.4	Iris-setosa
6	7	4.6	3.4	1.4	0.3	Iris-setosa
7	8	5.0	3.4	1.5	0.2	Iris-setosa

Latihan (12)

Tampilkan data pada 3 (tiga) baris terakhir dari dataset, dengan dataframe.

In [106…

```
#Latihan (12)
#Tampilkan data pada 3 (tiga) baris terakhir dari dataset, dengan dataframe

df.tail(3)
```

Out[106…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

Deskripsi Statistik Data

DataFrame method describe() menampilkan statistik dasar setiap kolom data yang bertipe numerik, mencakup banyaknya data (count), rerata aritmetik (mean), simpangan baku (std), nilai terkecil (min), kuartil pertama (25%), kuartil kedua/median (50%), kuartil ketiga (75%), dan nilai terbesar (max).

Latihan (13)

Hitung korelasi dari dataset. Dengan menggunakan method function

In [107…

```
#Latihan (13)
#Hitung korelasi dataset

df.corr()
```

Out[107…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
Id	1.000000	0.716676	-0.397729	0.882747	0.899759
SepalLengthCm	0.716676	1.000000	-0.109369	0.871754	0.817954
SepalWidthCm	-0.397729	-0.109369	1.000000	-0.420516	-0.356544
PetalLengthCm	0.882747	0.871754	-0.420516	1.000000	0.962757
PetalWidthCm	0.899759	0.817954	-0.356544	0.962757	1.000000

Latihan (14)

Berdasarkan pada perhitungan korelasi di Latihan (11), apakah yang dapat Bapak/Ibu simpulkan sementara? Silakan tuliskan simpulan sementara Bapak/Ibu pada cell di bawah ini.

In [108…

```
#latihan (14)
#Simpulan Sementara Hasil Korelasi di latihan (13)

# Terdapat banyak kolom yang berkorelasi positif satu sama lain
# maupun berkorelasi negatif seperti pada kolom PetalLengthCm - PetalWidthCm yang
# memiliki korelasi positif tertinggi sehingga memiliki hubungan yang kuat antar keduanya.
```

Latihan (15)

Hitung korelasi untuk kolom berikut ini: PetalLengthCm, PetalWidthCm

In [109…

```
#Latihan (15)
#Hitung korelasi dataset untuk kolom PetalLengthCm, PetalWidthCm

np.corrcoef(df['PetalLengthCm'], df['PetalWidthCm'])[0, 1]
```

Out[109…

```
0.9627570970509659
```

Latihan (16)

Method "describe" secara otomatis melakukan komputasi statistik untuk semua continous variable. Secara default "describe" melakukan ignore terhadap variabel bertype objek.

Komputasi statistik yang dilakukan terdiri dari: count, mean, std, min, max, 25%, 75%, max.

Latihan: Gunakan method describe pada dataset yang sudah di load untuk semua continous variabel. (Dataset Iris.csv)

In [110…

```
#Latihan (16)
# Penggunaan Metode describe untuk komputasi statistik

df.describe()
```

Out[110…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

Latihan (17)

Gunakan method describe pada dataset yang sudah di load untuk data bertype objek. (Dataset Iris.csv)

In [111…

```
#Latihan (17)
#Gunakan method describe pada dataset yang sudah di load untuk data bertype objek

df.describe(include=['O'])
```

Out[111…

	Species
count	150
unique	3
top	Iris-setosa
freq	50

Latihan 18

Gunakan method describe pada dataset yang sudah di load untuk semua type data (continous variabel dan type object).

In [112…

```
#Latihan (18)
#Gunakan method describe pada dataset yang sudah di load untuk semua type data

df.describe(include='all')
```

Out[112…

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
count	150.000000	150.000000	150.000000	150.000000	150.000000	150
unique	NaN	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	NaN	Iris-setosa
freq	NaN	NaN	NaN	NaN	NaN	50
mean	75.500000	5.843333	3.054000	3.758667	1.198667	NaN
std	43.445368	0.828066	0.433594	1.764420	0.763161	NaN
min	1.000000	4.300000	2.000000	1.000000	0.100000	NaN
25%	38.250000	5.100000	2.800000	1.600000	0.300000	NaN
50%	75.500000	5.800000	3.000000	4.350000	1.300000	NaN
75%	112.750000	6.400000	3.300000	5.100000	1.800000	NaN
max	150.000000	7.900000	4.400000	6.900000	2.500000	NaN

Latihan (19)

Hitunglah nilai mean dari dataset.

In [113…

```
#Latihan (19)
#Hitung nilai Mean dari dataset

df.mean()
```

Out[113…

```
Id                75.500000
SepalLengthCm      5.843333
SepalWidthCm       3.054000
PetalLengthCm      3.758667
PetalWidthCm       1.198667
dtype: float64
```

Latihan (20)

Hitung nilai mean dari dataset untuk kolom PetalLengthCm.

In [114…

```
#Latihan (20)
#Hitung nilai Mean untuk kolom PetalLengthCm

df['PetalLengthCm'].mean()
```

Out[114…

```
3.758666666666666
```

Latihan (21)

Carilah nilai minimal dari dataset untuk kolom SepalWidthCm.

In [115…

```
#Latihan (21)
#Cari nilai minimal untuk kolom SepalWidthCm

min(df['SepalWidthCm'])
```

Out[115…

```
2.0
```

Method Groupby

Method groupby memungkinkan analisis dilakukan secara per kelompok nilai atribut tertentu.

Latihan (22)

Hitunglah nilai mean dari dataset untuk kolom SepalLengthCm per Species dengan menggunakan metode groupby.

In [116…

```
#Latihan (22)
#Hitung nilai mean dari dataset untuk SepalLengthCm per Species dengan metode groupby

df.groupby('Species')['SepalLengthCm'].mean()
```

Out[116…

```
Species
Iris-setosa        5.006
Iris-versicolor    5.936
Iris-virginica     6.588
Name: SepalLengthCm, dtype: float64
```

Method Value Count

value_counts() menghasilkan frekuensi setiap nilai unik di dalam kolom, dan yang tertinggi count-nya adalah merupakan modus pada kolom tersebut.

Latihan (23)

Hitunglah frekuensi pada kolom 'Species' dengan menggunakan metode value_counts().

In [117…

```
#Latihan (23)
#Hitung frekuensi pada kolom 'Species' dengan menggunakan metode value_counts()

df['Species'].value_counts()
```

Out[117…

```
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: Species, dtype: int64
```

Latihan (24)

Tampilkan perhitungan frekuensi pada kolom 'Species' dengan menggunakan metode value_counts() dalam bentuk dataframe.

In [118…

```
#Latihan (24)
#Perhitungan frekuensi pada kolom 'Species' dengan menggunakan metode value_counts() dalam bentuk dataframe

pd.DataFrame(df['Species'].value_counts())
```

Out[118…

	Species
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Latihan (25)

Hitunglah frekuensi pada kolom 'PetalLenghCm' dengan menggunakan metode value_counts() dan dalam bentuk dataframe.

In [119…

```
#Latihan (25)
# Hitung frekuensi pada kolom 'PetalLenghCm' dengan menggunakan metode value_counts()

pd.DataFrame(df['PetalLengthCm'].value_counts())
```

Out[119…

	PetalLengthCm
1.5	14
1.4	12
5.1	8
4.5	8
1.6	7
1.3	7
5.6	6
4.0	5
4.9	5
4.7	5
5.0	4
1.7	4
4.8	4
4.4	4
4.2	4
4.1	3
5.7	3
5.5	3
6.1	3
3.9	3
4.6	3
5.8	3
5.2	2
1.9	2
6.0	2
1.2	2
4.3	2
5.3	2
5.4	2
3.3	2
6.7	2
3.5	2
5.9	2
3.6	1
3.8	1
1.0	1
3.0	1
6.3	1
6.6	1
3.7	1
1.1	1
6.4	1
6.9	1