



Hands-On

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

Pertemuan 5

Pertemuan 5 (Ilma) 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 [17]: 1 # Install modul kaggle secara inline (di dalam notebook)
2 pip install kaggle
```

Requirement already satisfied: kaggle in c:\anaconda\envs\new\lib\site-packages (1.5.12)
Requirement already satisfied: urllib3 in c:\anaconda\envs\new\lib\site-packages (from kaggle) (1.26.4)
Requirement already satisfied: python-dateutil in c:\anaconda\envs\new\lib\site-packages (from kaggle) (2.8.1)
Requirement already satisfied: requests in c:\anaconda\envs\new\lib\site-packages (from kaggle) (2.25.1)
Requirement already satisfied: python-slugify in c:\anaconda\envs\new\lib\site-packages (from kaggle) (5.0.2)
Requirement already satisfied: tqdm in c:\anaconda\envs\new\lib\site-packages (from kaggle) (4.66.0)
Requirement already satisfied: six>=1.10 in c:\users\ilham\appdata\roaming\python\python38\site-packages (from kaggle) (1.14.0)
Requirement already satisfied: certifi in c:\users\ilham\appdata\roaming\python\python38\site-packages (from kaggle) (2019.11.2)
Requirement already satisfied: text-unidecode>=1.3 in c:\anaconda\envs\new\lib\site-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: idna<3,>=2.5 in c:\anaconda\envs\new\lib\site-packages (from requests->kaggle) (2.10)
Requirement already satisfied: chardet<5,>=3.0.2 in c:\anaconda\envs\new\lib\site-packages (from requests->kaggle) (4.0.0)

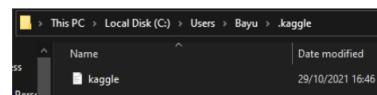
```
In [18]: 1 # Install modul kaggle secara eksternal melalui anaconda prompt:
```

The screenshot shows the Anaconda Prompt window with the command `pip install kaggle` entered and its output displayed. The output shows various package requirements and dependencies being installed.

2. Create Token API kaggle:

The screenshot shows the Kaggle website's account page. The 'Account' tab is selected. Under the 'API' section, it says: "Using Kaggle's beta API, you can interact with Competitions and Datasets to download data, make submissions, and more. Read the docs". There are two buttons: "Create New API Token" and "Expire API Token".

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<Windows-username>\kaggle\ (Windows)

Jika folder tersebut belum ada:

1. Buat folder di direktori C:\Users<Windows-username>\kaggle\
2. letakkan file kaggle.json kedalam folder tersebut

3. Download Dataset dari Kaggle:

```
usage: kaggle datasets [-h]
                      [list,files,download,create,version,init,metadata,status] ...
optional arguments:
  -h, --help            show this help message and exit
commands:
```

```
(list, files, download, create, version, init, metadata, status)
list      List available datasets
files     List dataset files
download  Download dataset files
create    Create a new dataset
version   Create or update dataset version
init     Initialize metadata file for dataset creation
metadata Download metadata about a dataset
status   Get the creation status for a dataset
```

Dokumentasi Kaggle Commands selengkapnya [Disini](#)

In [22]:

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

ref	title	size	lastUpdated
downloadCount			
voteCount			
usabilityRating			
--	--	--	--
uciml/iris	Iris Species	4KB	2016-09-27 07:38:
05	226326 2680 0.7941176		
anshid/iris-flower-dataset	Iris Flower Dataset	1010B	2018-03-22 15:18:
06	40606 370 0.8235294		
vikrishnan/iris-dataset	Iris Dataset	999B	2017-08-03 16:00:
44	2930 26 0.7647059		
therohk/ireland-historical-news	Irish Times - Waxy-Wany News	52MB	2021-09-25 10:52:
48	2983 157 1.0		
chuckyin/iris-datasets	Iris datasets	1KB	2017-03-10 09:35:
43	1771 14 0.7352941		
rmatman/iris-dataset-json-version	Iris Dataset (JSON Version)	1KB	2018-04-06 20:21:
31	5634 43 0.75		
parulpandey/palmer-archipelago-antarctica-penguin-data	Palmer Archipelago (Antarctica) penguin data	11KB	2020-06-09 10:14:
54	10042 114 0.9705882		
conorrot/irish-weather-hourly-data	Irish Weather (hourly data)	67MB	2020-06-29 20:15:
18	1864 40 0.8235294		
saurabh00007/iriscsv	Iris.csv	1KB	2017-11-09 07:34:
35	17139 57 0.4117647		
jillanisoftech/iris-dataset-uci	Iris dataset uci	1KB	2021-11-06 15:11:
47	37 12 1.0		
fleanend/birds-songs-numeric-dataset	Birds' Songs Numeric Dataset	25MB	2019-04-01 09:09:
46	706 25 0.9411765		
kamrankausar/iris-data	iris_data	1KB	2017-11-30 10:26:
01	1117 13 0.64705884		
jeffheaton/iris-computer-vision	Iris Computer Vision	5MB	2020-11-24 21:23:
29	306 9 0.875		
styven/iris-dataset	Iris dataset	1KB	2017-11-04 14:10:
12	793 8 0.29411766		
arslanali4343/iris-species	Iris Species	2KB	2020-07-02 06:09:
09	61 13 0.5625		
olgabelitskaya/flower-color-images	Flower Color Images	50MB	2020-10-01 22:48:
07	8363 161 0.75		
naureenmohammad/mm-iris-dataset	MMU iris dataset	30MB	2020-07-25 18:38:
33	645 19 0.5625		
rutujavaidya/iris-dataset	Iris Dataset	1KB	2021-07-25 17:37:
14	36 6 0.4117647		
shantanuss/iris-flower-dataset	IRIS flower dataset	1KB	2020-01-18 19:43:
18	197 3 0.9411765		
ashishs0ni/iris-dataset	Iris_dataset	1KB	2018-08-05 14:26:
19	600 7 0.64705884		

In [23]:

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

Downloading iris.zip to C:\Users\Ilham\Desktop\Kuliah\Spda Diktii\Tugas 5

```
0%|          | 0.00/3.60k [00:00<?, ?B/s]
100%#####| 3.60k/3.60k [00:00<00:00, 3.78MB/s]
```

Atau bisa juga menggunakan link dari kaggle

Latihan (1)

Silahkan Download sebuah dataset menggunakan API Kaggle

In [5]:

```
1 #Latihan (1)
2 #Langkah nya seperti contoh diatas
```

PENGUNAAN 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 [2]:

```
1 #Latihan(2)
2 #Import Library Pandas
3
4 import pandas as pd
5
6 #Import Library Numpy
7
8 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 syntax: 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 [3]:

```
1 #Latihan(3)
2 #Panggil file (Load file bernama Iris.csv) dan simpan dalam dataframe lalu tampilkan 5 baris awal dataset dengan function he
3
4 df = pd.read_csv('iris.csv')
5 df.head()
```

Out[3]:

	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 [4]: 1 #Latihan(4)
          2 #Tampilkan tipe data dari kolom yang ada pada dataset
          3
          4 df.dtypes
```

```
Out[4]: 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 [5]: 1 #Latihan (5)
          2 #Tipe Data dari kolom yang ada di dataset
          3
          4 #Kolom "Id" memiliki tipe data = int64
          5 #Kolom "SepalLengthCm" memiliki tipe data = float64
          6 #Kolom "Species" memiliki tipe data = object
```

Latihan (6)

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

```
In [6]: 1 #Latihan (6)
          2 #Hitung ukuran (jumlah baris dan kolom) dari dataset
          3
          4 df.shape
```

```
Out[6]: (150, 6)
```

Latihan (7)

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

```
In [7]: 1 #Latihan (7)
          2
          3 #Jumlah Baris pada dataset adalah = 150
          4
          5 #Jumlah kolom pada dataset adalah = 6
```

Latihan (8) ¶

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

```
In [8]: 1 #Latihan (8)
          2 #Tampilkan data untuk kolom "Id" dan kolom "Species" dalam bentuk dataframe
          3
          4 df[["Id", "Species"]]
```

```
Out[8]:   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 [9]: 1 #Latihan (9)
          2 #Tampilkan data dengan dataframe, dan data yang ditampilkan adalah baris dengan indeks 0 (nol) sampai dengan indeks 9 (sembilan)
          3
          4 df.head(9)
```

```
Out[9]:   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
```

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 [10]: 1 #Latihan (10)
          2 #Tampilkan data hanya kolom "Id" dan kolom "Species", pada baris dengan indeks 0 (nol) sampai dengan indeks 9 (sembilan)
          3
          4 df[["Id", "Species"]].head(9)
```

```
Out[10]:
```

	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

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 [11]:
```

```
1 #Latihan (11)
2 #Tampilkan data pada 8 (delapan) baris pertama dari dataset, dengan dataframe
3
4 df.head(8)
```

```
Out[11]:
```

	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 [12]:
```

```
1 #Latihan (12)
2 #Tampilkan data pada 3 (tiga) baris terakhir dari dataset, dengan dataframe
3
4 df.tail(3)
```

```
Out[12]:
```

	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 [13]:
```

```
1 #Latihan (13)
2 #Hitung korelasi dataset
3
4 df.corr()
```

```
Out[13]:
```

	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

```
In [14]:
```

```
1 df[["PetalLengthCm", "PetalWidthCm"]].corr()
```

```
Out[14]:
```

	PetalLengthCm	PetalWidthCm
PetalLengthCm	1.000000	0.962757
PetalWidthCm	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 [16]:
```

```
1 #Latihan (14)
2 #Simpulan Sementara Hasil Korelasi di Latihan (13)
3
4 #Panjang dari sepal tidak mempengaruhi lebar dari sepal, tetapi sangat sangat berpengaruh pada Panjang dan lebar petal.
5 #panjang dan lebar petal sangat berkaitan satu sama Lain
```

Latihan (15)

Hitung korelasi untuk kolom berikut ini: PetalLengthCm, PetalWidthCm

```
In [28]:
```

```
1 #Latihan (15)
2 #Hitung korelasi dataset untuk kolom PetalLengthCm, PetalWidthCm
3
4 df[["PetalLengthCm", "PetalWidthCm"]].corr()
```

```
Out[28]:
```

	PetalLengthCm	PetalWidthCm
PetalLengthCm	1.000000	0.962757
PetalWidthCm	0.962757	1.000000

Latihan (16)

Method "describe" secara otomatis melakukan komputasi statistik untuk semua continuous 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 continuous variabel. (Dataset Iris.csv)

```
In [34]: 1 #Latihan (16)
2 # Penggunaan Metode describe untuk komputasi statistik
3
4 df.describe(include=np.number)

Out[34]:   Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
count 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 bertipe objek. (Dataset Iris.csv)

```
In [35]: 1 #Latihan (17)
2 #Gunakan method describe pada dataset yang sudah di Load untuk data bertipe objek
3
4 df.describe(include=np.object)

Out[35]:   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 (continuous variabel dan type object).

```
In [37]: 1 #Latihan (18)
2 #Gunakan method describe pada dataset yang sudah di load untuk semua type data
3
4 df.describe(include='all')

Out[37]:   Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species
count 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 [39]: 1 #Latihan (19)
2 #Hitung nilai Mean dari dataset
3
4 df.mean()

Out[39]: Id      75.50000
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 [40]: 1 #Latihan (20)
2 #Hitung nilai Mean untuk kolom PetalLengthCm
3
4 df['PetalLengthCm'].mean()

Out[40]: 3.7586666666666693
```

Latihan (21)

Carilah nilai minimal dari dataset untuk kolom SepalWidthCm.

```
In [41]: 1 #Latihan (21)
2 #Cari nilai minimal untuk kolom SepalWidthCm
3
4 df['SepalWidthCm'].min()

Out[41]: 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 [44]: 1 #Latihan (22)
2 #Hitung nilai mean dari dataset untuk SepalLengthCm per Species dengan metode groupby
3
4 df.groupby(['Species']).mean()[['SepalLengthCm']]

Out[44]: SepalLengthCm
Species
```

Iris-setosa	5 006
Iris-versicolor	5 936
Iris-virginica	6 588

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 [47]: 1 #Latihan (23)
2 #Hitung frekuensi pada kolom 'Species' dengan menggunakan metode value_counts()
3
4 df['Species'].value_counts()

Out[47]: 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 [49]: 1 #Latihan (24)
2 #Perhitungan frekuensi pada kolom 'Species' dengan menggunakan metode value_counts() dalam bentuk dataframe
3
4 tempDF = df['Species'].value_counts().rename_axis('unique_values').reset_index(name='counts')
5 tempDF

Out[49]:   unique_values  counts
0      Iris-setosa       50
1    Iris-versicolor      50
2   Iris-virginica       50
```

Latihan (25)

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

```
In [53]: 1 #Latihan (25)
2 # Hitung frekuensi pada kolom 'PetalLenghCm' dengan menggunakan metode value_counts()
3
4 tempDF = df['PetalLengthCm'].value_counts().rename_axis('unique_values').reset_index(name='counts')
5 tempDF

Out[53]:   unique_values  counts
0            1.5       14
1            1.4       12
2            5.1        8
3            4.5        8
4            1.3        7
5            1.6        7
6            5.6        6
7            4.0        5
8            4.9        5
9            4.7        5
10           4.8        4
11           1.7        4
12           4.4        4
13           4.2        4
14           5.0        4
15           4.1        3
16           5.5        3
17           4.6        3
18           6.1        3
19           5.7        3
20           3.9        3
21           5.8        3
22           1.2        2
23           1.9        2
24           6.7        2
25           3.5        2
26           5.9        2
27           6.0        2
28           5.4        2
29           5.3        2
30           3.3        2
31           4.3        2
32           5.2        2
33           6.3        1
34           1.1        1
35           6.4        1
36           3.6        1
37           3.7        1
38           3.0        1
39           3.8        1
40           6.6        1
41           6.9        1
42           1.0        1
```