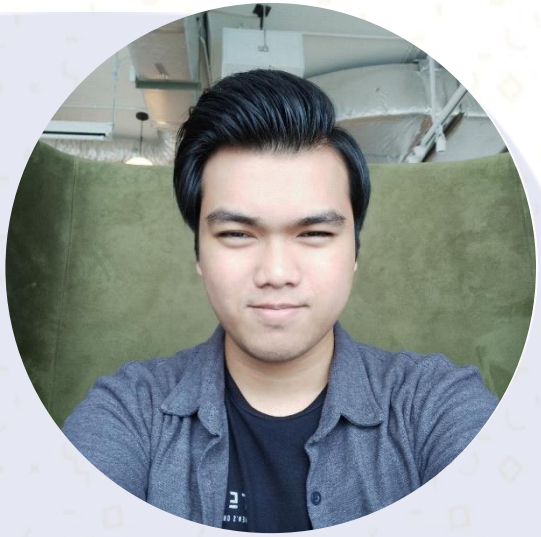
Intermediate Dataframe

Profile



linkedin.com/in/romansyasetyo/

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

1. Sorting in Dataframe
2. Filtering Dataframe
3. Creating Additional Column
4. Grouping & Aggregate in Dataframe
5. Combining Dataframe

Hands on using Python

DigitalSkola
Uncover The World Of Digital Skills With Us

Before we start.....

Google Colab

https://colab.research.google.com/

Review

What is Dataframe?

col1	col2	col3	Col4

Now, we will learn how to manipulate dataframe

Very similar with manipulating data in SQL,

but here we will use Python

Sorting in Dataframe

Sorting in Dataframe

```
>>> df
   col1  col2  col3 col4
0     A     2     0    a
1     A     1     1    B
2     B     9     9    c
3   NaN     8     4    D
4     D     7     2    e
5     C     4     3    F
```

Sort → Mengurutkan

- Berdasarkan abjad
- Berdasarkan angka
- Dsb,

Sorting in Dataframe

```
>>> df
   col1  col2  col3 col4
0     A     2     0    a
1     A     1     1    B
2     B     9     9    c
3   NaN     8     4    D
4     D     7     2    e
5     C     4     3    F
```

DataFrame.**sort_values**(by, axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last', ignore_index=False, key=None)

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html

Sorting in Dataframe

Case 1 : Mengurutkan berdasarkan abjad di col1 – part 1

```
>>> df
   col1  col2  col3 col4
0     A     2     0    a
1     A     1     1    B
2     B     9     9    c
3   NaN     8     4    D
4     D     7     2    e
5     C     4     3    F
```

```
>>> df.sort_values(by=['col1'])
   col1  col2  col3 col4
0     A     2     0    a
1     A     1     1    B
2     B     9     9    c
5     C     4     3    F
4     D     7     2    e
3   NaN     8     4    D
```

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html

Sorting in Dataframe

Case 2 : Mengurutkan berdasarkan abjad di col1 – part 2

```
>>> df
   col1  col2  col3 col4
0     A     2     0    a
1     A     1     1    B
2     B     9     9    c
3   NaN     8     4    D
4     D     7     2    e
5     C     4     3    F
```

```
>>> df.sort_values(by='col1', ascending=False)
   col1  col2  col3 col4
4     D     7     2    e
5     C     4     3    F
2     B     9     9    c
0     A     2     0    a
1     A     1     1    B
3   NaN     8     4    D
```

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html

Sorting in Dataframe

Case 3 : Mengurutkan berdasarkan abjad di col1 – part 3

```
>>> df
   col1  col2  col3 col4
0     A     2     0    a
1     A     1     1    B
2     B     9     9    c
3   NaN     8     4    D
4     D     7     2    e
5     C     4     3    F
```

```
>>> df.sort_values(by='col1', ascending=False, na_position='first')
   col1  col2  col3 col4
3   NaN     8     4    D
4     D     7     2    e
5     C     4     3    F
2     B     9     9    c
0     A     2     0    a
1     A     1     1    B
```

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html

Sorting in Dataframe

Case 4 : Mengurutkan berdasarkan lebih dari 1 column

```
>>> df
   col1  col2  col3 col4
0     A     2     0    a
1     A     1     1    B
2     B     9     9    c
3   NaN     8     4    D
4     D     7     2    e
5     C     4     3    F
```

```
>>> df.sort_values(by=['col1', 'col2'])
   col1  col2  col3 col4
1     A     1     1    B
0     A     2     0    a
2     B     9     9    c
5     C     4     3    F
4     D     7     2    e
3   NaN     8     4    D
```

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html

Filtering Dataframe

Filtering Dataframe

Melakukan seleksi terhadap dataframe untuk mendapatkan hanya informasi yang dibutuhkan/diinginkan

- Get only some column
- Filter by single condition
- Filter by multiple condition

	sepal_length	sepal_width	petal_length	petal_width	flower_class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Filtering Dataframe

- Get only some columns

```
df[['sepal_length','sepal_width']]
```

	sepal_length	sepal_width
0	5.1	3.5
1	4.9	3.0
2	4.7	3.2
3	4.6	3.1
4	5.0	3.6

```
df.filter(items=['sepal_length', 'sepal_width'])
```

	sepal_length	sepal_width
0	5.1	3.5
1	4.9	3.0
2	4.7	3.2
3	4.6	3.1
4	5.0	3.6

Filtering Dataframe

- Get some columns with filter using loc & iloc

loc : **label**-based, perlu specify nama column & row

iloc : integer **index**-based, perlu specify index dari column & row

Filtering Dataframe

- Get some columns with filter using loc & iloc

```
df.loc[df.flower_class == 'Iris-setosa']
```

	sepal_length	sepal_width	petal_length	petal_width	flower_class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Filtering Dataframe

- Get some columns with filter using loc & iloc

```
df.loc[(df.flower_class == 'Iris-setosa') & (df.sepal_length == 5.1)]
```

	sepal_length	sepal_width	petal_length	petal_width	flower_class
0	5.1	3.5	1.4	0.2	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa

Filtering Dataframe

- Get some columns with filter using loc & iloc

```
[12] df.loc[(df.flower_class == 'Iris-setosa') & (df.sepal_length == 5.1), ['flower_class', 'sepal_length']]
```

	flower_class	sepal_length
0	Iris-setosa	5.1
17	Iris-setosa	5.1
19	Iris-setosa	5.1
21	Iris-setosa	5.1

Filtering Dataframe

- Get some columns with filter using loc & iloc

```
[16] df.iloc[[0,3]]
```

	sepal_length	sepal_width	petal_length	petal_width	flower_class
0	5.1	3.5	1.4	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa

Filtering Dataframe

- Get some columns with filter using loc & iloc

```
[19] df.iloc[[0,2],[1,3]]
```

	sepal_width	petal_width
0	3.5	0.2
2	3.2	0.2

Filtering Dataframe

- Get some columns with filter using loc & iloc

```
[18] df.iloc[0:3]
```

	sepal_length	sepal_width	petal_length	petal_width	flower_class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa

Filtering Dataframe

- Get some columns with filter using loc & iloc

```
df.iloc[1:3, 2:4]
```

	petal_length	petal_width
1	1.4	0.2
2	1.3	0.2

Creating Additional Column

Creating Additional Column

Memberikan column tambahan pada dataframe

- Column tambahan : value dari column yang lain
- Column tambahan : single value
- Column tambahan : other

Creating Additional Column

- Column tambahan : value dari column yang lain

```
df['sepal_length_v2'] = df['sepal_length'] * 100
df
```

	sepal_length	sepal_width	petal_length	petal_width	flower_class	sepal_length_v2
0	5.1	3.5	1.4	0.2	Iris-setosa	510.0
1	4.9	3.0	1.4	0.2	Iris-setosa	490.0
2	4.7	3.2	1.3	0.2	Iris-setosa	470.0
3	4.6	3.1	1.5	0.2	Iris-setosa	460.0
4	5.0	3.6	1.4	0.2	Iris-setosa	500.0

DigitalSkola
Uncover The World Of Digital Skills With Us

Grouping & Aggregate in Dataframe

Grouping & Aggregate in Dataframe

Melakukan grouping dan aggregate, pada dasarnya bertujuan untuk mendapat summary atau rangkuman dari dataframe

- Menggunakan groupby
- Menggunakan pivot_table

Grouping & Aggregate in Dataframe

- Menggunakan groupby

```
df.groupby('flower_class').count()
```

flower_class	sepal_length	sepal_width	petal_length	petal_width
Iris-setosa	50	50	50	50
Iris-versicolor	50	50	50	50
Iris-virginica	50	50	50	50

Grouping & Aggregate in Dataframe

- Menggunakan groupby

```
df.groupby('flower_class')['sepal_length'].mean()
```

```
flower_class
Iris-setosa        5.006
Iris-versicolor    5.936
Iris-virginica     6.588
Name: sepal_length, dtype: float64
```

Grouping & Aggregate in Dataframe

- Menggunakan groupby

```
df.groupby('flower_class').agg( average_per_class = ('sepal_length', 'mean'),
                                median_per_class = ('sepal_length', 'median') )
```

flower_class	average_per_class	median_per_class
Iris-setosa	5.006	5.0
Iris-versicolor	5.936	5.9
Iris-virginica	6.588	6.5

Grouping & Aggregate in Dataframe

- Menggunakan pivot_table

pandas.**pivot_table**(data, values=None, index=None, columns=None, aggfunc='mean', fill_value=None, margins=False, dropna=True, margins_name='All', observed=False)

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.pivot_table.html

Grouping & Aggregate in Dataframe

- Menggunakan pivot_table

```
pd.pivot_table(df, values = 'sepal_length', index = 'flower_class', aggfunc = ['mean', 'median'])
```

	mean	median
	sepal_length	sepal_length
flower_class		
Iris-setosa	5.006	5.0
Iris-versicolor	5.936	5.9
Iris-virginica	6.588	6.5

Combining Dataframe

Combining Dataframe

Melakukan proses kombinasi dataframe, tujuan utamanya untuk melengkapi atau memperkaya data yang ada

- Melakukan concatenate
- Melakukan merge

Combining Dataframe

- Melakukan concatenate

```
pd.concat([df1, df2, df3])
```

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df2

	A	B	C	D
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7

df3

	A	B	C	D
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

Result

		A	B	C	D
x	0	A0	B0	C0	D0
x	1	A1	B1	C1	D1
x	2	A2	B2	C2	D2
x	3	A3	B3	C3	D3
y	4	A4	B4	C4	D4
y	5	A5	B5	C5	D5
y	6	A6	B6	C6	D6
y	7	A7	B7	C7	D7
z	8	A8	B8	C8	D8
z	9	A9	B9	C9	D9
z	10	A10	B10	C10	D10
z	11	A11	B11	C11	D11

Combining Dataframe

- Melakukan merge

DataFrame.**merge**(right, how='inner', on=None, left_on=None, right_on=None, left_index=False, right_index=False, sort=False, suffixes=('_x', '_y'), copy=True, indicator=False, validate=None)

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.merge.html

Combining Dataframe

- Melakukan merge

```
>>> df1
    lkey value
0    foo     1
1    bar     2
2    baz     3
3    foo     5
>>> df2
    rkey value
0    foo     5
1    bar     6
2    baz     7
3    foo     8
```

```
>>> df1.merge(df2, left_on='lkey', right_on='rkey')
  lkey  value_x rkey  value_y
0  foo        1  foo        5
1  foo        1  foo        8
2  foo        5  foo        5
3  foo        5  foo        8
4  bar        2  bar        6
5  baz        3  baz        7
```

DigitalSkola
Uncover The World Of Digital Skills With Us

Keep practicing

in your Google Colab!

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank You