Pandas Dataframe - Advanced

Table of Content

What will We Learn Today?

1. Indexing Dataframe
2. Menghapus Variable/kolom
3. Menggabungkan Dataframe
4. Concatenate & Append Dataframe
5. Pivot Table Dataframe
6. Melting Dataframe
7. Fungsi Lambda dalam Dataframe

The Almighty Pandas

Indexing Dataframe

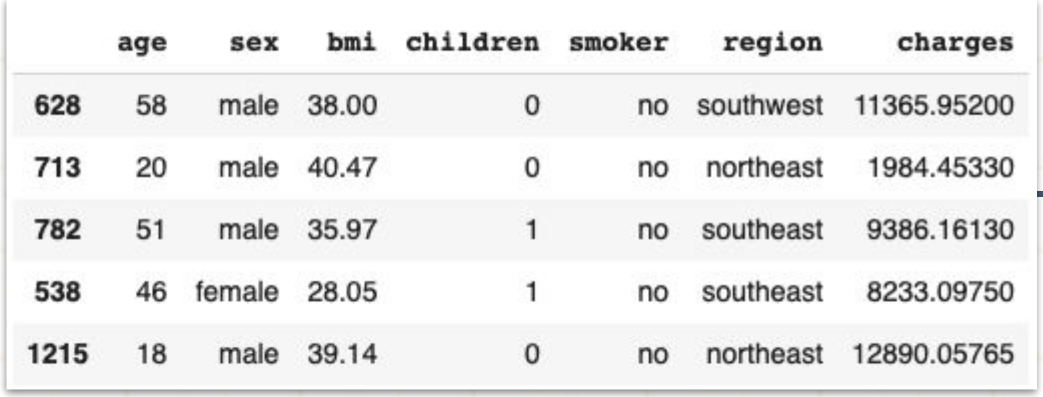
Indexing pada dataframe menggunakan Pandas memiliki beberapa pengaplikasian di dalam dataset.

Contoh:

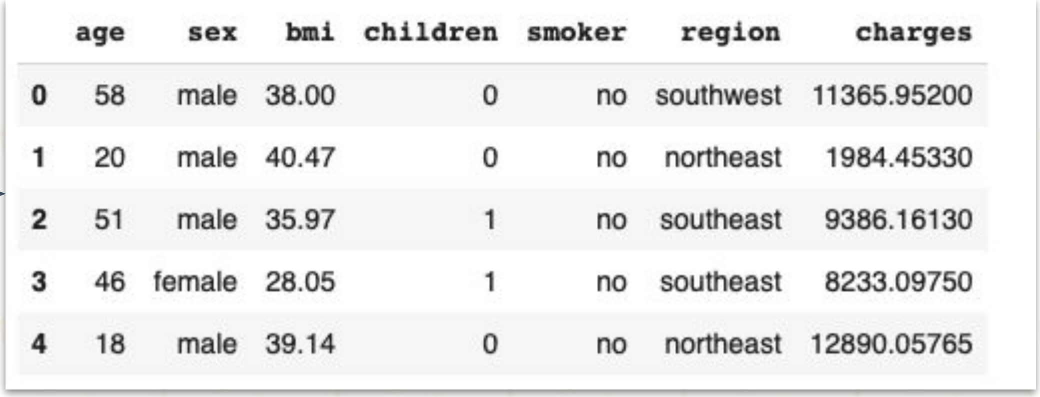
- *Mengurutkan index*
- *Membuat data pada variable tertentu menjadi index*

Indexing Dataframe *(Reset Index)*

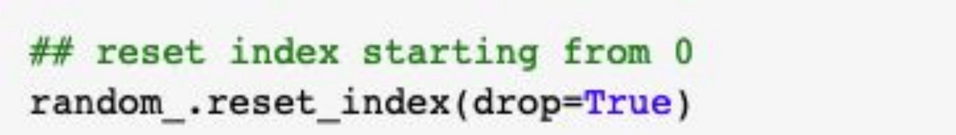
Index sangat membantu dalam mencari data ketika ingin melakukan kalkulasi terdapat data di dalam dataset.



	age	sex	bmi	children	smoker	region	charges
628	58	male	38.00	0	no	southwest	11365.95200
713	20	male	40.47	0	no	northeast	1984.45330
782	51	male	35.97	1	no	southeast	9386.16130
538	46	female	28.05	1	no	southeast	8233.09750
1215	18	male	39.14	0	no	northeast	12890.05765



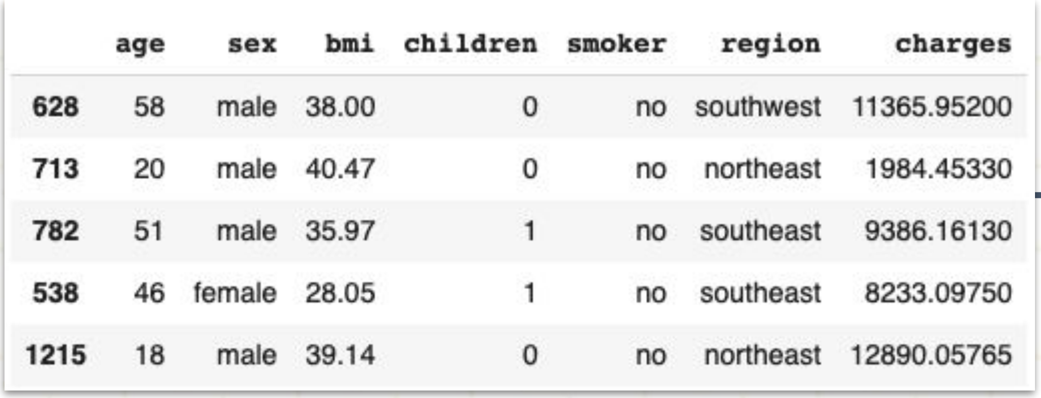
	age	sex	bmi	children	smoker	region	charges
0	58	male	38.00	0	no	southwest	11365.95200
1	20	male	40.47	0	no	northeast	1984.45330
2	51	male	35.97	1	no	southeast	9386.16130
3	46	female	28.05	1	no	southeast	8233.09750
4	18	male	39.14	0	no	northeast	12890.05765



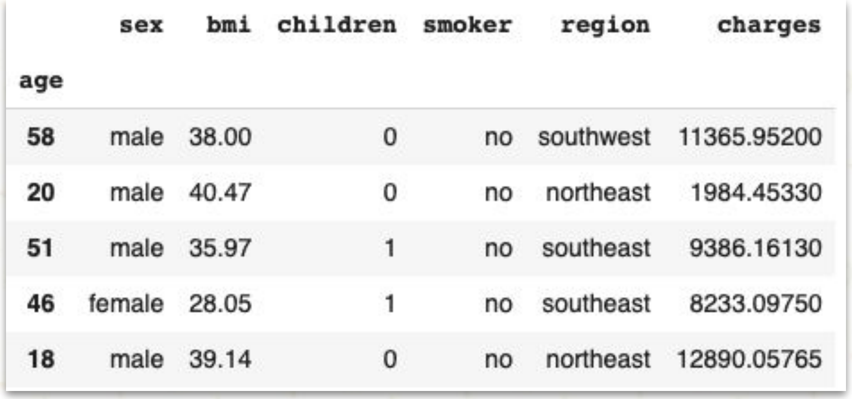
```
## reset index starting from 0
random_.reset_index(drop=True)
```

Indexing Dataframe *(Set Column as Index)*

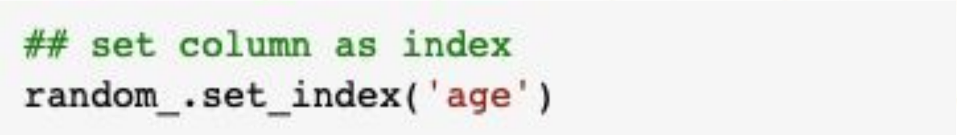
Index juga dapat membuat index sendiri berdasarkan dari kolom yang ada di dalam dataset



	age	sex	bmi	children	smoker	region	charges
628	58	male	38.00	0	no	southwest	11365.95200
713	20	male	40.47	0	no	northeast	1984.45330
782	51	male	35.97	1	no	southeast	9386.16130
538	46	female	28.05	1	no	southeast	8233.09750
1215	18	male	39.14	0	no	northeast	12890.05765



	sex	bmi	children	smoker	region	charges
age						
58	male	38.00	0	no	southwest	11365.95200
20	male	40.47	0	no	northeast	1984.45330
51	male	35.97	1	no	southeast	9386.16130
46	female	28.05	1	no	southeast	8233.09750
18	male	39.14	0	no	northeast	12890.05765



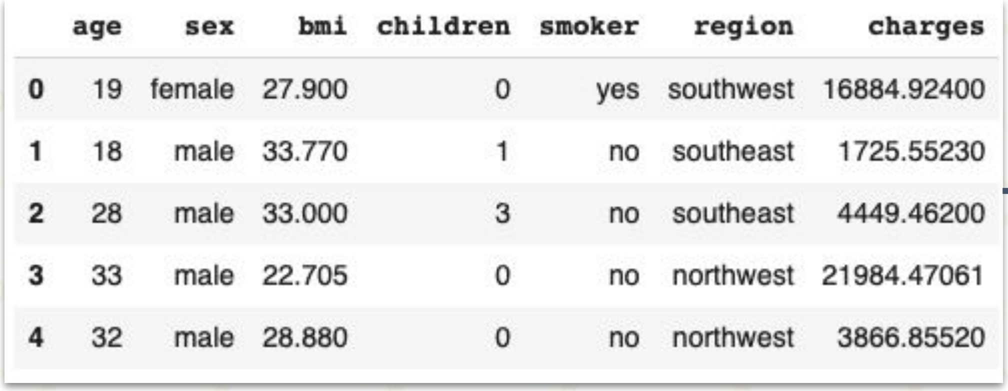
```
## set column as index
random_.set_index('age')
```

Menghapus Variable/Kolom

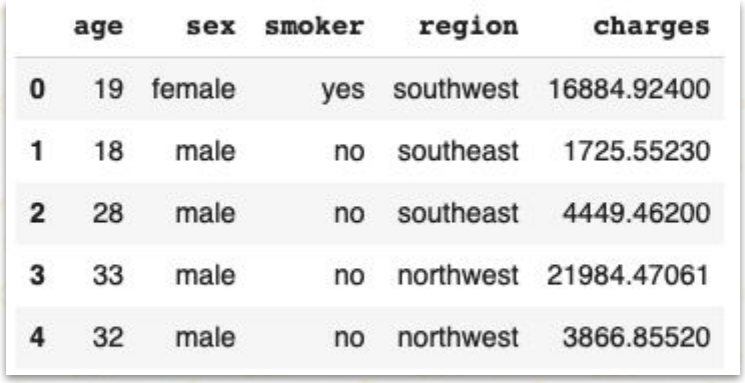
Pandas dapat menghapus kolom-kolom yang tidak diinginkan. Adapun tujuan menghapus kolom adalah:

- *Untuk memilih kolom yang akan dianalisa*
- *Untuk memilih kolom yang digunakan dalam machine learning model*

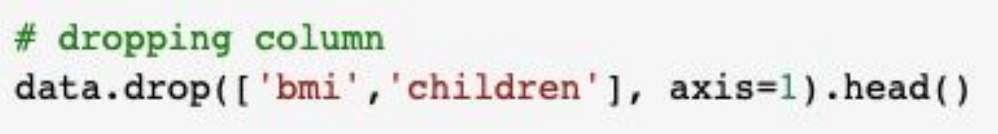
Dropping Column *(beberapa kolom)*



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520



	age	sex	smoker	region	charges
0	19	female	yes	southwest	16884.92400
1	18	male	no	southeast	1725.55230
2	28	male	no	southeast	4449.46200
3	33	male	no	northwest	21984.47061
4	32	male	no	northwest	3866.85520



```
# dropping column
data.drop(['bmi','children'], axis=1).head()
```

Menggabungkan Dataframe

*Dataset juga dapat digabungkan selain menggunakan metode merge, yaitu metode **JOIN**.*

Join Dataframe

Selain menggunakan merge, pandas juga dapat menggabungkan dua dataset menjadi satu menggunakan **join**. Terdapat perbedaan antara merge dan join yaitu:

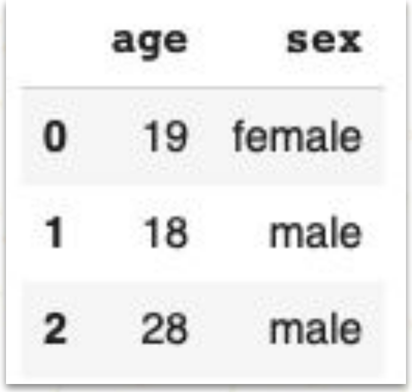
→ Join

- Menggabungkan data berdasarkan index

→ Merge

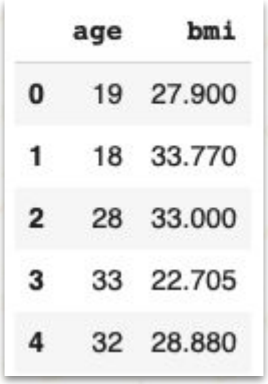
- Menggabungkan data lebih fleksibel dan memungkinkan untuk menentukan kolom selain index untuk kedua dataframe

Join Dataframe



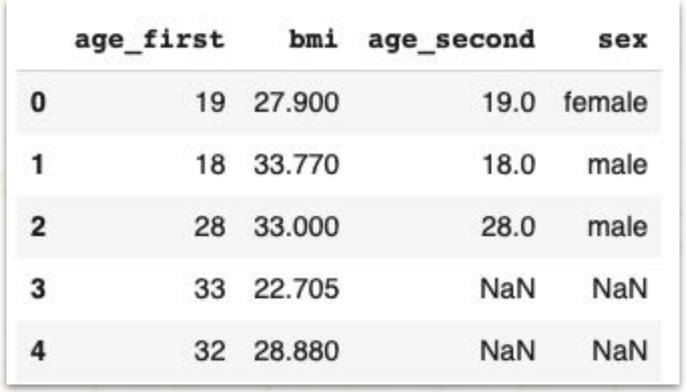
	age	sex
0	19	female
1	18	male
2	28	male

+

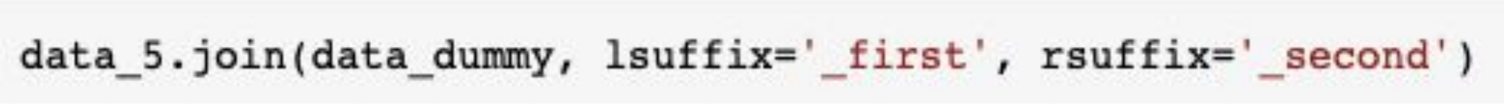


	age	bmi
0	19	27.900
1	18	33.770
2	28	33.000
3	33	22.705
4	32	28.880

→



	age_first	bmi	age_second	sex
0	19	27.900	19.0	female
1	18	33.770	18.0	male
2	28	33.000	28.0	male
3	33	22.705	NaN	NaN
4	32	28.880	NaN	NaN

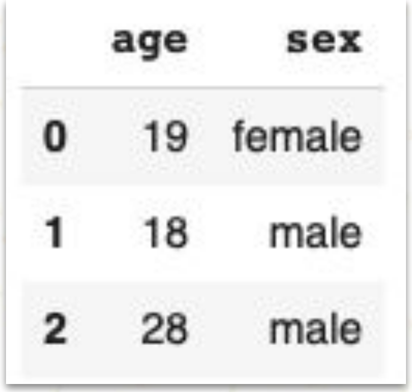


```
data_5.join(data_dummy, lsuffix='_first', rsuffix='_second')
```

Concatenate & Append Dataframe

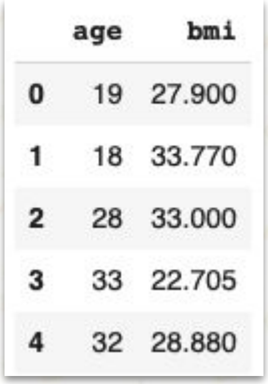
Menggabungkan objek dengan Pandas pada spesifik axis baik itu x-axis (horizontal) ataupun y-axis (vertikal)

Concatenate *(Horizontal)*



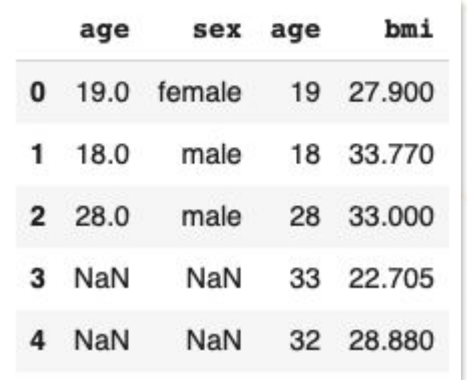
	age	sex
0	19	female
1	18	male
2	28	male

+

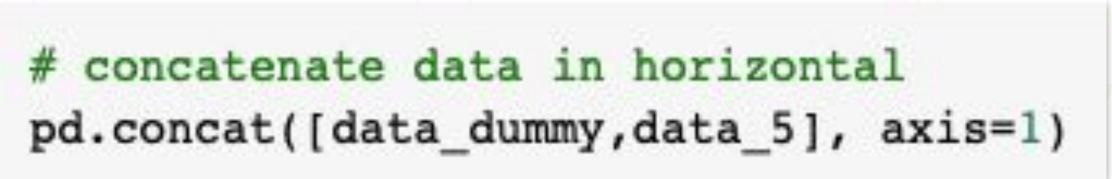


	age	bmi
0	19	27.900
1	18	33.770
2	28	33.000
3	33	22.705
4	32	28.880

→

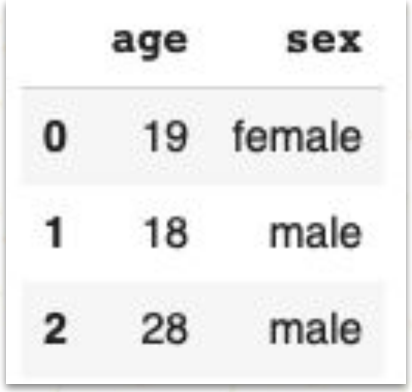


	age	sex	age	bmi
0	19.0	female	19	27.900
1	18.0	male	18	33.770
2	28.0	male	28	33.000
3	NaN	NaN	33	22.705
4	NaN	NaN	32	28.880



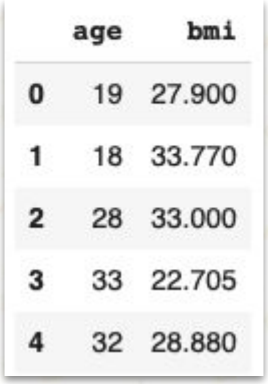
```
# concatenate data in horizontal
pd.concat([data_dummy,data_5], axis=1)
```

Concatenate *(Vertical)*

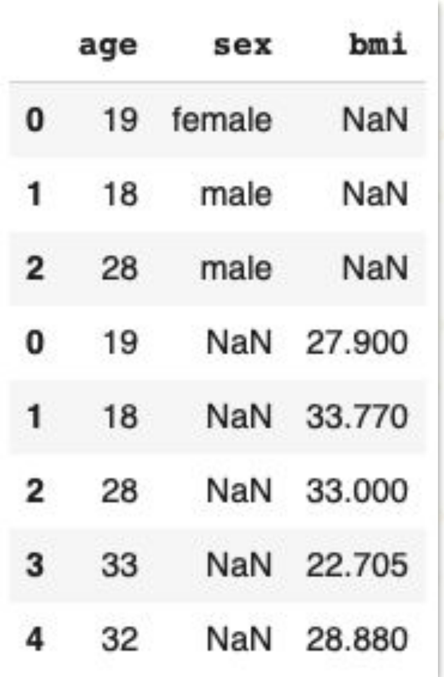


	age	sex
0	19	female
1	18	male
2	28	male

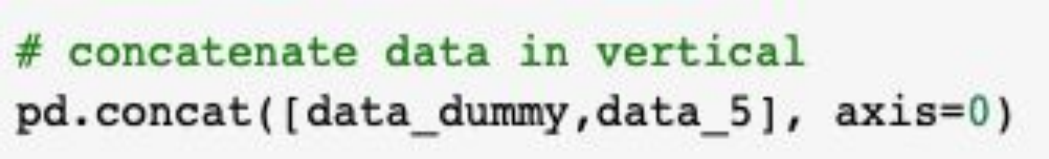
+



	age	bmi
0	19	27.900
1	18	33.770
2	28	33.000
3	33	22.705
4	32	28.880



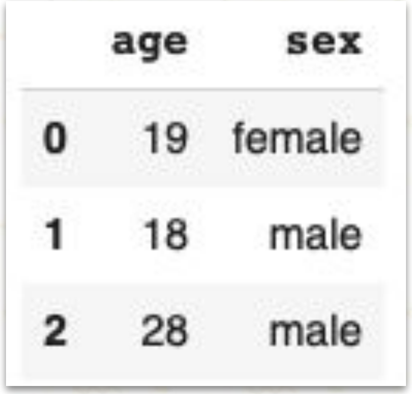
	age	sex	bmi
0	19	female	NaN
1	18	male	NaN
2	28	male	NaN
0	19	NaN	27.900
1	18	NaN	33.770
2	28	NaN	33.000
3	33	NaN	22.705
4	32	NaN	28.880



```
# concatenate data in vertical
pd.concat([data_dummy,data_5], axis=0)
```

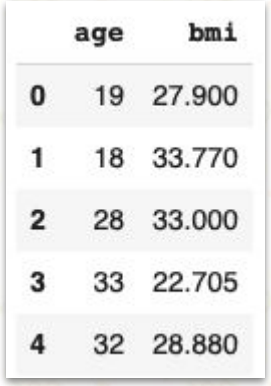
Append

Dalam dataframe, append dapat dilakukan jika terdapat nama kolom pada kedua dataset yang sama

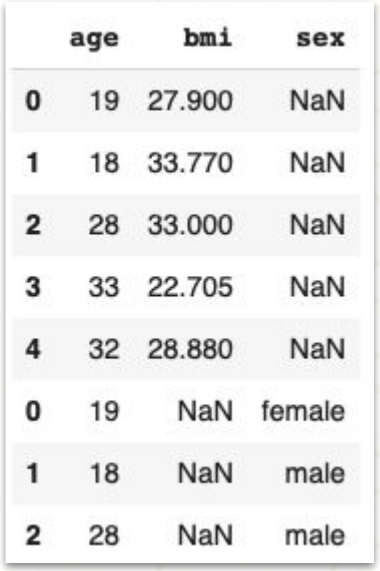


	age	sex
0	19	female
1	18	male
2	28	male

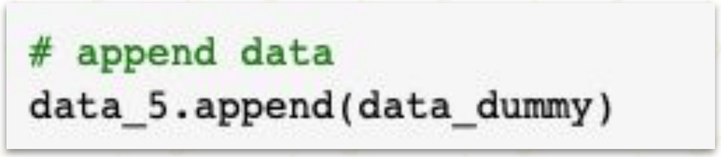
+



	age	bmi
0	19	27.900
1	18	33.770
2	28	33.000
3	33	22.705
4	32	28.880



	age	bmi	sex
0	19	27.900	NaN
1	18	33.770	NaN
2	28	33.000	NaN
3	33	22.705	NaN
4	32	28.880	NaN
0	19	NaN	female
1	18	NaN	male
2	28	NaN	male



```
# append data
data_5.append(data_dummy)
```

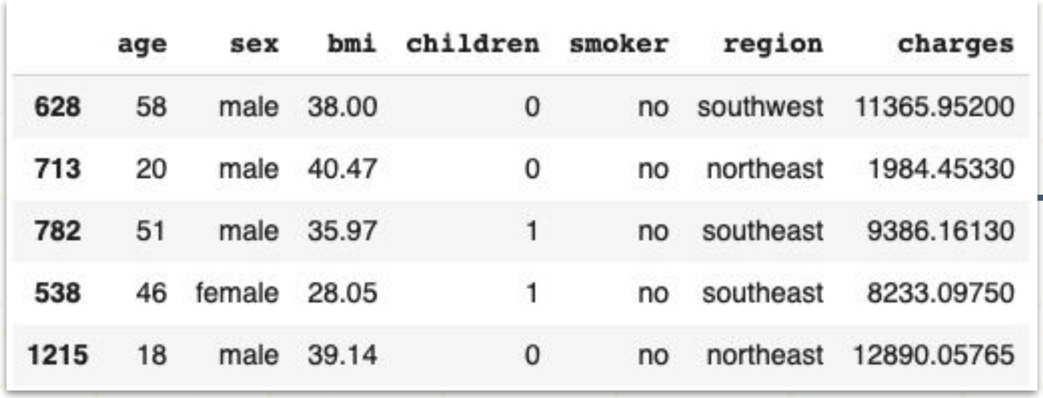
Pivot Table Dataframe

Pivot table memberikan informasi berupa agregasi suatu data dengan melampirkan isi data pada nama kolom tertentu

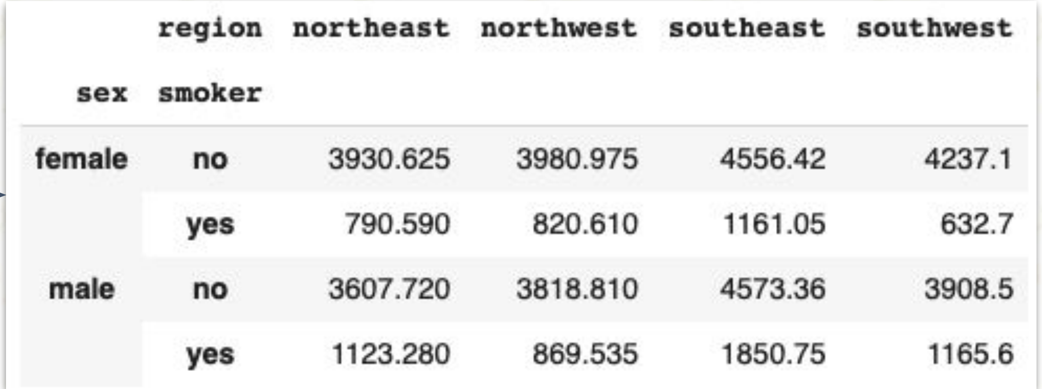
Pivot Table

Beberapa karakteristik pivot table menggunakan pandas:

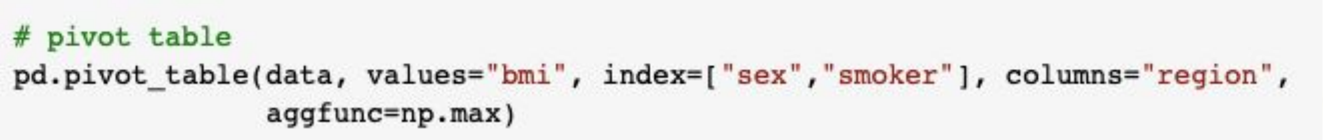
- Tampilan seperti pivot table yang ada di spreadsheet
- Nama kolom sebagai level data disimpan dalam bentuk MultiIndex



	age	sex	bmi	children	smoker	region	charges
628	58	male	38.00	0	no	southwest	11365.95200
713	20	male	40.47	0	no	northeast	1984.45330
782	51	male	35.97	1	no	southeast	9386.16130
538	46	female	28.05	1	no	southeast	8233.09750
1215	18	male	39.14	0	no	northeast	12890.05765



	region	northeast	northwest	southeast	southwest
sex	smoker				
female	**no**	3930.625	3980.975	4556.42	4237.1
	yes	790.590	820.610	1161.05	632.7
male	**no**	3607.720	3818.810	4573.36	3908.5
	yes	1123.280	869.535	1850.75	1165.6

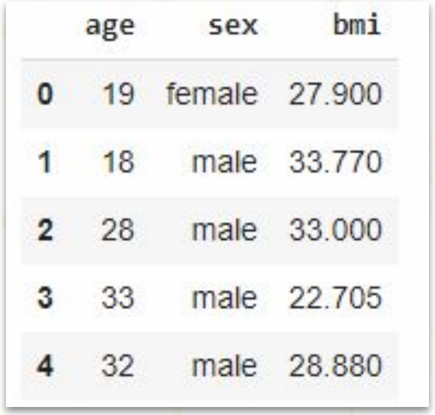


```
# pivot table
pd.pivot_table(data, values="bmi", index=["sex","smoker"], columns="region",
               aggfunc=np.max)
```

Melting Dataframe

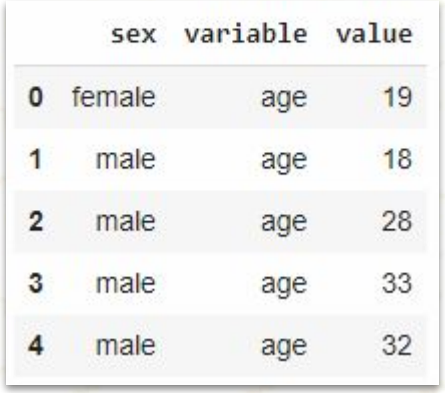
*Melting dataframe digunakan untuk memberikan informasi data dimana **nama kolom/variable akan menjadi datapoint** dan tetap memberikan informasi nilai dari kolom/variable namun di kolom yang berbeda*

Pivot Table

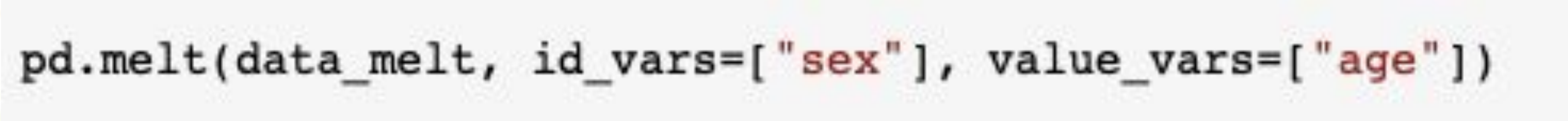


	age	sex	bmi
0	19	female	27.900
1	18	male	33.770
2	28	male	33.000
3	33	male	22.705
4	32	male	28.880

→



	sex	variable	value
0	female	age	19
1	male	age	18
2	male	age	28
3	male	age	33
4	male	age	32



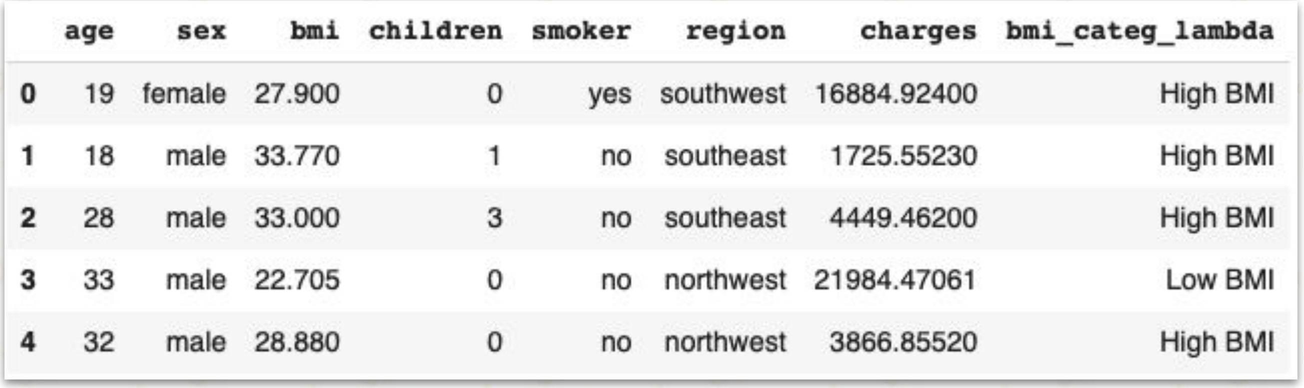
```
pd.melt(data_melt, id_vars=["sex"], value_vars=["age"])
```

DigitalSkola
Uncover The World Of Digital Skills With Us

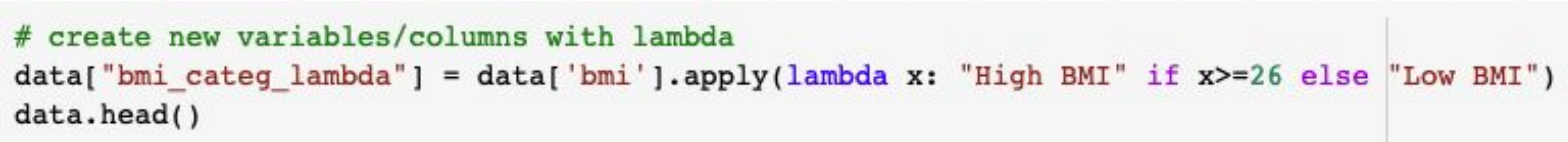
Lambda Function

Lambda function mempersingkat syntax python

Lambda Function

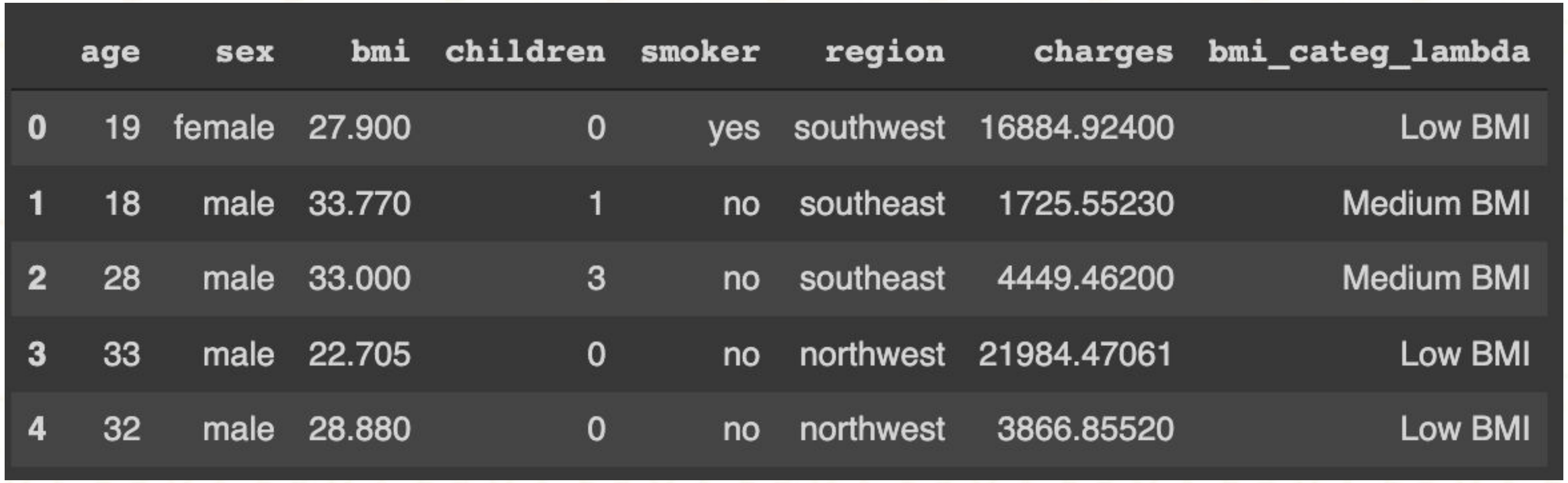


	age	sex	bmi	children	smoker	region	charges	bmi_categ_lambda
0	19	female	27.900	0	yes	southwest	16884.92400	High BMI
1	18	male	33.770	1	no	southeast	1725.55230	High BMI
2	28	male	33.000	3	no	southeast	4449.46200	High BMI
3	33	male	22.705	0	no	northwest	21984.47061	Low BMI
4	32	male	28.880	0	no	northwest	3866.85520	High BMI



```
# create new variables/columns with lambda
data["bmi_categ_lambda"] = data['bmi'].apply(lambda x: "High BMI" if x>=26 else "Low BMI")
data.head()
```

Homework



	age	sex	bmi	children	smoker	region	charges	bmi_categ_lambda
0	19	female	27.900	0	yes	southwest	16884.92400	Low BMI
1	18	male	33.770	1	no	southeast	1725.55230	Medium BMI
2	28	male	33.000	3	no	southeast	4449.46200	Medium BMI
3	33	male	22.705	0	no	northwest	21984.47061	Low BMI
4	32	male	28.880	0	no	northwest	3866.85520	Low BMI

Hitung:

- min, max, dan mean dari kolom bmi_categ_lambda menggunakan pivot table
- Ubah data point di kolom sex, region dan bmi_categ_lambda menjadi huruf besar semua menggunakan lambda function

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU