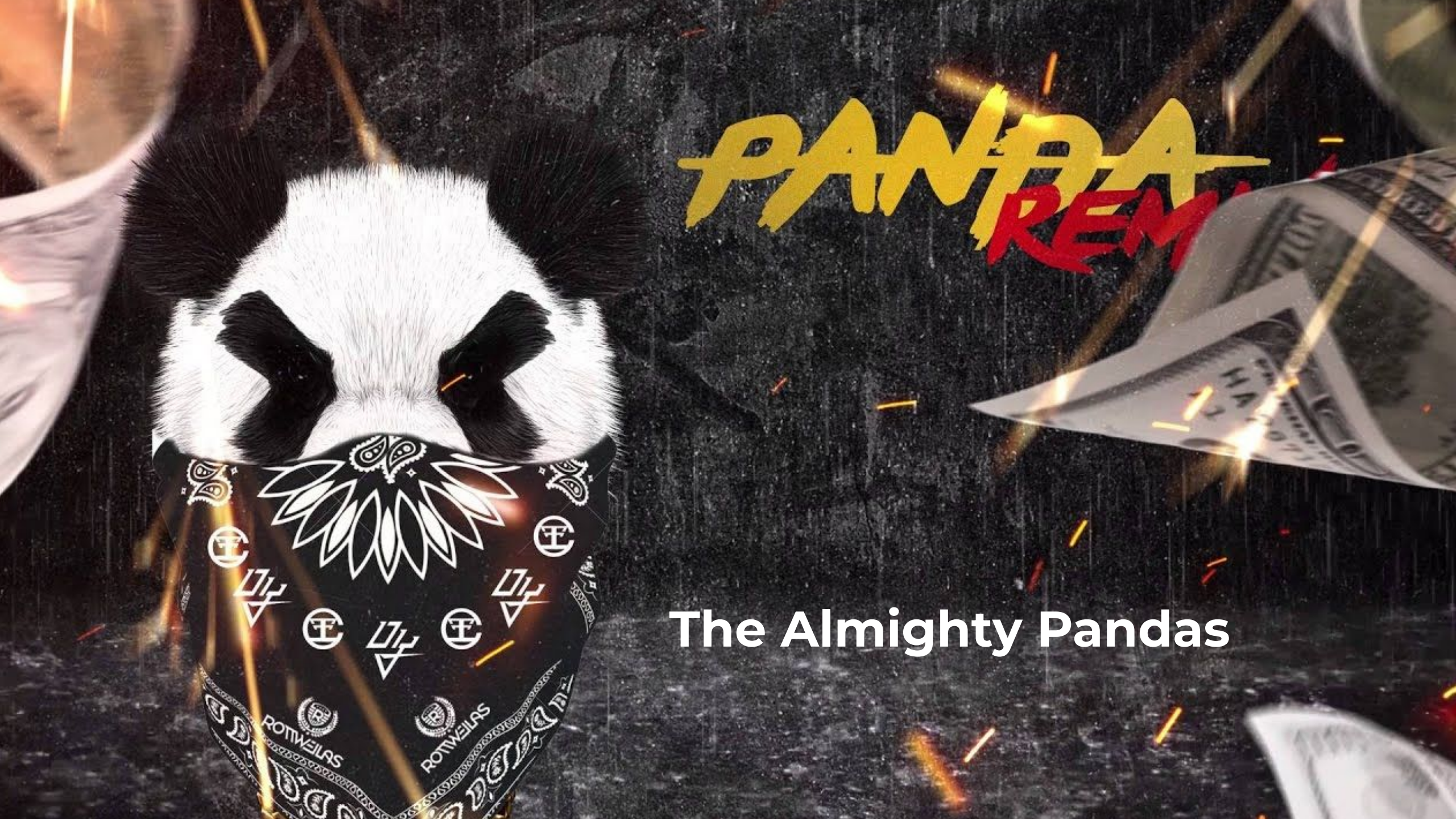
Pandas Dataframe - Intermediate

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

1. Sorting Dataframe
2. Filtering Dataframe
3. Membuat Kolom tambahan
4. Mengelompokkan Data
5. Menggabungkan Data antar Dataframe



The Almighty Pandas

Sorting

Pandas dapat melakukan pengurutan suatu data berdasarkan perintah dan output yang diinginkan.

Contoh:

- *Mengurutkan dari yang terkecil*
- *Mengurutkan sesuai alphabet*

Sorting in Dataframe *(Alphabetic)*

	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	bmi	children	smoker	region	charges
0	19	female	27.90	0	yes	southwest	16884.9240
714	24	female	22.60	0	no	southwest	2457.5020
716	49	female	22.61	1	no	northwest	9566.9909
718	51	female	36.67	2	no	northwest	10848.1343
719	58	female	33.44	0	no	northwest	12231.6136

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

```
## Alphabetic sorting
data.sort_values(by=['sex'])
```

DigitalSkola
Uncover The World Of Digital Skills With Us

Sorting in Dataframe *(Alphabetic - Descending)*

	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	bmi	children	smoker	region	charges
446	60	male	29.64	0	no	northeast	12730.9996
1052	49	male	29.83	1	no	northeast	9288.0267
1070	37	male	37.07	1	yes	southeast	39871.7043
550	63	male	30.80	0	no	southwest	13390.5590
1068	63	male	21.66	1	no	northwest	14349.8544

```
## Alphabetic sorting descending
data.sort_values(by=['sex'], ascending=False)
```

DigitalSkola
Uncover The World Of Digital Skills With Us

Sorting in Dataframe *(Numerical)*

	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	bmi	children	smoker	region	charges
1248	18	female	39.82	0	no	southeast	1633.96180
482	18	female	31.35	0	no	southeast	1622.18850
492	18	female	25.08	0	no	northeast	2196.47320
525	18	female	33.88	0	no	southeast	11482.63485
529	18	male	25.46	0	no	northeast	1708.00140

```
## Numerical sorting ascending
data.sort_values(by=['age'])
```

DigitalSkola
Uncover The World Of Digital Skills With Us

Sorting in Dataframe *(2 kolom sekaligus)*

	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	bmi	children	smoker	region	charges
172	18	male	15.960	0	no	northeast	1694.79640
250	18	male	17.290	2	yes	northeast	12829.45510
359	18	female	20.790	0	no	southeast	1607.51010
1212	18	male	21.470	0	no	northeast	1702.45530
1033	18	male	21.565	0	yes	northeast	13747.87235

```
## Sorting more than two columns
data.sort_values(by=['age','bmi'])
```

Filtering

Pandas dapat menyeleksi beberapa data berdasarkan perintah dan output yang diinginkan. Contoh:

- *Seleksi kolom tertentu*
- *Seleksi baris dan kolom*
- *Seleksi data tertentu pada suatu kolom*

DigitalSkola
Uncover The World Of Digital Skills With Us

Filtering in Dataframe *(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
0	19	female
1	18	male
2	28	male
3	33	male
4	32	male

```
data[["age","sex"]]
```

```
data.filter(items=["age","sex"])
```

Filtering in Dataframe

Pada Dataframe, terdapat 2 fungsi untuk melakukan seleksi

- **loc** untuk melakukan seleksi kolom dan baris yang berdasar pada **nama** baris dan kolom
- **iloc** untuk melakukan seleksi kolom dan baris yang berdasar **indeks** baris dan kolom

Filtering in Dataframe *(loc)*

	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	bmi	children	smoker	region	charges
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
8	37	male	29.830	2	no	northeast	6406.41070

```
# loc
data.loc[(data["sex"] == "male")]
```

DigitalSkola
Uncover The World Of Digital Skills With Us

Filtering in Dataframe *(iloc)*

	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	bmi	children	smoker	region	charges
10	25	male	26.22	0	no	northeast	2721.3208
11	62	female	26.29	0	yes	southeast	27808.7251
12	23	male	34.40	0	no	southwest	1826.8430
13	56	female	39.82	0	no	southeast	11090.7178
14	27	male	42.13	0	yes	southeast	39611.7577

```
# iloc with index 10 to 20
data.iloc[10:21,:]
```

Creating new variable

Pandas juga dapat melakukan penambahan variable dari dataset yang ada. Contoh:

- *Melakukan standar kalkulasi (+, -, x, :)*
- *Melakukan pengelompokan suatu data*

DigitalSkola
Uncover The World Of Digital Skills With Us

Creating New Variable

	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	bmi	children	smoker	region	charges	discount_charges
0	19	female	27.900	0	yes	southwest	16884.92400	2532.738600
1	18	male	33.770	1	no	southeast	1725.55230	258.832845
2	28	male	33.000	3	no	southeast	4449.46200	667.419300
3	33	male	22.705	0	no	northwest	21984.47061	3297.670591
4	32	male	28.880	0	no	northwest	3866.85520	580.028280

```
data['discount_charges'] = data['charges']*0.15
data.head()
```

Grouping

Pandas memiliki fungsi untuk mengelompokkan data yang ada dan melakukan perhitungan dalam waktu yang sama. Contoh:

- *Melakukan pengelompokan suatu data*

Grouping

	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

sex	average_age	median_charges
female	39.503021	9412.96250
male	38.917160	9369.61575

```
# grouping average age and median charges in sex variable
data.groupby('sex').agg(average_age = ('age', 'mean'),
                        median_charges = ('charges','median'))
```

Merging Dataframe

Pandas memiliki fungsi untuk menggabungkan dataset berdasarkan kemiripan suatu fitur

DigitalSkola
Uncover The World Of Digital Skills With Us

Merging

	age	sex	bmi	children	smoker	region	charges	discount_charges	Name
0	19	female	27.900	0	yes	southwest	16884.92400	2532.738600	doni
1	18	male	33.770	1	no	southeast	1725.55230	258.832845	jojo
2	28	male	33.000	3	no	southeast	4449.46200	667.419300	beki
3	33	male	22.705	0	no	northwest	21984.47061	3297.670591	madrid
4	32	male	28.880	0	no	northwest	3866.85520	580.028280	milan

+

	Type of House	Price House	Name
0	House 1	10000	doni
1	House 2	12300	jojo
2	House 3	11000	beki
3	House 4	14000	madrid
4	House 5	12100	milan

```
data_5.merge(data2, how='inner', on='Name')
```

	age	sex	bmi	children	smoker	region	charges	discount_charges	Name	Type of House	Price House
0	19	female	27.900	0	yes	southwest	16884.92400	2532.738600	doni	House 1	10000
1	18	male	33.770	1	no	southeast	1725.55230	258.832845	jojo	House 2	12300
2	28	male	33.000	3	no	southeast	4449.46200	667.419300	beki	House 3	11000
3	33	male	22.705	0	no	northwest	21984.47061	3297.670591	madrid	House 4	14000
4	32	male	28.880	0	no	northwest	3866.85520	580.028280	milan	House 5	12100

Homework

Terdapat tiga jenis dataset:

1. data.csv
2. data_2.csv
3. final.csv (hasil gabungan dari dataset pertama dan kedua)

Tugas anda adalah menggabungkan secara penuh (full outer join) dari dataset pertama (data.csv) dengan dataset kedua (data_2.csv). Tugas tambahannya adalah:

1. Ambil dataset pertama yang tidak match (clue-nya adalah **awards_won_y** yang null)
2. Cek apakah dimensi hasil dari pertanyaan pertama yang dilakukan sama dengan dataset final.csv
3. Ambil kolom yang merupakan bagian dataset pertama saja
4. Samakan nama kolom dari pertanyaan ketiga dengan nama kolom dari dataset data.csv

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU