WEEK 6 - HOMEWORK 3

Intermediate Dataframe I

By: Anugrah Yazid Ghani

HOMEWORK DESCRIPTION

WEEK 6 - HOMEWORK 3

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

```
[98] # Import Library
     import pandas as pd
     import numpy as np
```

DigitalSkola

DATA 1

```
# Data Pertama
data1 = pd.read_csv('https://raw.githubusercontent.com/densaiko/data_science_learning/main/dataset/data.csv')
data1.sort_values(by=['employee_id'])
```

	Unnamed: 0	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won	avg_training_score	is_promoted
52690	52690	1	Analytics	region_7	Bachelor's	m	sourcing	2	29	3.0	5	0	85.0	0
10257	10257	2	Finance	region_2	Master's & above	f	sourcing	1	35	1.0	2	0	63.0	0
32895	32895	4	Sales & Marketing	region_2	Bachelor's	m	other	1	25	3.0	2	0	53.0	0
4424	4424	5	Analytics	region_7	Master's & above	m	other	2	46	3.0	7	0	86.0	0
41261	41261	7	Operations	region_32	Bachelor's	m	other	1	31	3.0	7	0	59.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
31061	31061	78292	Operations	region_2	Master's & above	m	sourcing	1	59	2.0	16	0	57.0	1
19088	19088	78294	Sales & Marketing	region_22	Bachelor's	m	sourcing	3	35	3.0	3	0	49.0	0
52714	52714	78296	Procurement	region_2	Bachelor's	f	sourcing	1	28	5.0	5	0	70.0	0
9030	9030	78297	Operations	region_13	Bachelor's	f	sourcing	1	34	5.0	7	0	56.0	0
52646	52646	78298	Procurement	region_2	Master's & above	f	referred	1	39	5.0	10	0	67.0	0

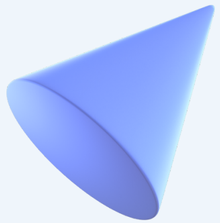
59808 rows × 14 columns

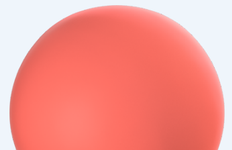
DATA 2

```
[106] # Data Kedua
      data2 = pd.read_csv('https://raw.githubusercontent.com/densaiko/data_science_learning/main/dataset/data_2.csv')
      data2.sort_values(by=['employee_id'])
```

	Unnamed: 0	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won	avg_training_score	is_promoted
4565	28461	18	Operations	region_24	Bachelor's	f	sourcing	1	33	3.0	4	1	57.0	0
4022	19135	26	Procurement	region_34	Bachelor's	m	sourcing	1	29	3.0	5	0	70.0	0
2713	2003	34	Operations	region_14	Bachelor's	m	other	1	43	3.0	13	0	NaN	0
1037	13597	61	Analytics	region_7	Master's & above	m	other	1	36	3.0	2	0	84.0	0
4718	41760	62	Sales & Marketing	region_2	Bachelor's	m	sourcing	1	39	3.0	8	0	50.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
811	3458	78287	Analytics	region_7	Bachelor's	m	other	1	35	3.0	2	0	83.0	0
4936	30366	78288	Sales & Marketing	region_32	Bachelor's	m	sourcing	1	29	3.0	3	1	60.0	1
3781	27484	78289	Sales & Marketing	region_6	Bachelor's	m	sourcing	2	30	NaN	1	0	52.0	0
3297	10949	78290	Operations	region_2	Bachelor's	f	other	1	52	1.0	7	0	60.0	0
1798	19088	78294	Sales & Marketing	region_22	Bachelor's	m	sourcing	3	35	3.0	3	0	49.0	0

5000 rows × 14 columns





DigitalSkola

DATA FINAL (1+2)

```
[107] # Data Final (hasil gabungan dari dataset pertama dan kedua)
      data_final = pd.read_csv('https://raw.githubusercontent.com/densaiko/data_science_learning/main/dataset/final.csv')
      data_final.sort_values(by=['employee_id'])
```

	Unnamed: 0	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won	avg_training_score	is_promoted
51943	57564	1	Analytics	region_7	Bachelor's	m	sourcing	2	29	3.0	5	0	85.0	0
10113	11191	2	Finance	region_2	Master's & above	f	sourcing	1	35	1.0	2	0	63.0	0
32482	35941	4	Sales & Marketing	region_2	Bachelor's	m	other	1	25	3.0	2	0	53.0	0
4379	4809	5	Analytics	region_7	Master's & above	m	other	2	46	3.0	7	0	86.0	0
40728	45094	7	Operations	region_32	Bachelor's	m	other	1	31	3.0	7	0	59.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34951	38699	78291	Sales & Marketing	region_2	Bachelor's	f	sourcing	1	29	3.0	3	0	50.0	0
30691	33930	78292	Operations	region_2	Master's & above	m	sourcing	1	59	2.0	16	0	57.0	1
51968	57590	78296	Procurement	region_2	Bachelor's	f	sourcing	1	28	5.0	5	0	70.0	0
8925	9855	78297	Operations	region_13	Bachelor's	f	sourcing	1	34	5.0	7	0	56.0	0
51904	57519	78298	Procurement	region_2	Master's & above	f	referred	1	39	5.0	10	0	67.0	0

54022 rows × 14 columns

DigitalSkola

FULL OUTER JOIN DATA 1 & 2

FULL OUTER JOIN DATA 1 & 2

```
[108] # Data Gabungan (Full Outer Join Dataset 1 & 2)
      data_baru = data1.merge(data2, how='outer', on='employee_id')
      data_baru.sort_values(by=['employee_id'])
```

	Unnamed: 0_x	employee_id	department_x	region_x	education_x	gender_x	recruitment_channel_x	no_of_trainings_x	age_x	previous_year_rating_x	length_of_service_x	awards_won_x	avg_training_score_x	i
57564	52690	1	Analytics	region_7	Bachelor's	m	sourcing	2	29	3.0	5	0	85.0	
11191	10257	2	Finance	region_2	Master's & above	f	sourcing	1	35	1.0	2	0	63.0	
35941	32895	4	Sales & Marketing	region_2	Bachelor's	m	other	1	25	3.0	2	0	53.0	
4809	4424	5	Analytics	region_7	Master's & above	m	other	2	46	3.0	7	0	86.0	
45094	41261	7	Operations	region_32	Bachelor's	m	other	1	31	3.0	7	0	59.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
33930	31061	78292	Operations	region_2	Master's & above	m	sourcing	1	59	2.0	16	0	57.0	
20809	19088	78294	Sales & Marketing	region_22	Bachelor's	m	sourcing	3	35	3.0	3	0	49.0	
57590	52714	78296	Procurement	region_2	Bachelor's	f	sourcing	1	28	5.0	5	0	70.0	
9855	9030	78297	Operations	region_13	Bachelor's	f	sourcing	1	34	5.0	7	0	56.0	
57519	52646	78298	Procurement	region_2	Master's & above	f	referred	1	39	5.0	10	0	67.0	

59858 rows × 27 columns

DigitalSkola

1. DATASET 1 THAT DON'T MATCH

1. DATASET 1 YANG TIDAK MATCH

```
[109] a = data_baru[data_baru['awards_won_y'].isnull()]
      data_baru_1 = a.iloc[:, :14]
      data_baru_1.sort_values(by=['employee_id'])
```

	Unnamed: 0_x	employee_id	department_x	region_x	education_x	gender_x	recruitment_channel_x	no_of_trainings_x	age_x	previous_year_rating_x	length_of_service_x	awards_won_x	avg_training_score_x	is_promoted_x
57564	52690	1	Analytics	region_7	Bachelor's	m	sourcing	2	29	3.0	5	0	85.0	0
11191	10257	2	Finance	region_2	Master's & above	f	sourcing	1	35	1.0	2	0	63.0	0
35941	32895	4	Sales & Marketing	region_2	Bachelor's	m	other	1	25	3.0	2	0	53.0	0
4809	4424	5	Analytics	region_7	Master's & above	m	other	2	46	3.0	7	0	86.0	0
45094	41261	7	Operations	region_32	Bachelor's	m	other	1	31	3.0	7	0	59.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
38699	35429	78291	Sales & Marketing	region_2	Bachelor's	f	sourcing	1	29	3.0	3	0	50.0	0
33930	31061	78292	Operations	region_2	Master's & above	m	sourcing	1	59	2.0	16	0	57.0	1
57590	52714	78296	Procurement	region_2	Bachelor's	f	sourcing	1	28	5.0	5	0	70.0	0
9855	9030	78297	Operations	region_13	Bachelor's	f	sourcing	1	34	5.0	7	0	56.0	0
57519	52646	78298	Procurement	region_2	Master's & above	f	referred	1	39	5.0	10	0	67.0	0

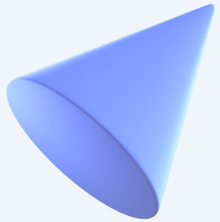
54022 rows × 14 columns

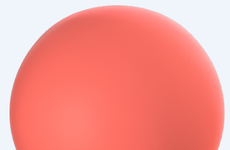
2. CHECK DIMENSION NO.1 WITH DIMENSION DATA FINAL

2. CHECK DIMENSI NO.1 DENGAN DIMENSI DATA FINAL

```
[111] # Pengecekkan Dimensi
      print("Dimensi Pernyataan No.1 :", data_baru_1.shape)
      print("Dimensi Data Final :", data_final.shape)

      Dimensi Pernyataan No.1 : (54022, 14)
      Dimensi Data Final : (54022, 14)
```





DigitalSkola

3. TAKE DATASET 1 COLUMNS

3. AMBIL KOLOM DATASET PERTAMA

```
[113] dataset_1 = data_baru.iloc[:, :14]
      dataset_1.sort_values(by=['employee_id'])
```

	Unnamed: 0_x	employee_id	department_x	region_x	education_x	gender_x	recruitment_channel_x	no_of_trainings_x	age_x	previous_year_rating_x	length_of_service_x	awards_won_x	avg_training_score_x	is_promoted_x
57564	52690	1	Analytics	region_7	Bachelor's	m	sourcing	2	29	3.0	5	0	85.0	0
11191	10257	2	Finance	region_2	Master's & above	f	sourcing	1	35	1.0	2	0	63.0	0
35941	32895	4	Sales & Marketing	region_2	Bachelor's	m	other	1	25	3.0	2	0	53.0	0
4809	4424	5	Analytics	region_7	Master's & above	m	other	2	46	3.0	7	0	86.0	0
45094	41261	7	Operations	region_32	Bachelor's	m	other	1	31	3.0	7	0	59.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
33930	31061	78292	Operations	region_2	Master's & above	m	sourcing	1	59	2.0	16	0	57.0	1
20809	19088	78294	Sales & Marketing	region_22	Bachelor's	m	sourcing	3	35	3.0	3	0	49.0	0
57590	52714	78296	Procurement	region_2	Bachelor's	f	sourcing	1	28	5.0	5	0	70.0	0
9855	9030	78297	Operations	region_13	Bachelor's	f	sourcing	1	34	5.0	7	0	56.0	0
57519	52646	78298	Procurement	region_2	Master's & above	f	referred	1	39	5.0	10	0	67.0	0

59858 rows × 14 columns

4. RENAME COLUMNS DATA NO.3 WITH COLUMNS DATA 1

4. RENAME KOLOM NO.3 DENGAN NAMA KOLOM DATA PERTAMA

```
# Rename Columns
rename_columns = dataset_1.rename({'Unnamed: 0_x': 'Unnamed: 0',
                                   'department_x': 'department',
                                   'region_x': 'region',
                                   'education_x': 'education',
                                   'gender_x': 'gender',
                                   'recruitment_channel_x': 'recruitment_channel',
                                   'no_of_trainings_x': 'no_of_trainings',
                                   'age_x': 'age',
                                   'previous_year_rating_x': 'previous_year_rating',
                                   'length_of_service_x': 'length_of_service',
                                   'awards_won_x': 'awards_won',
                                   'avg_training_score_x': 'avg_training_score',
                                   'is_promoted_x': 'is_promoted'},
                                  axis=1)

rename_columns.sort_values(by=['employee_id'])
```

	Unnamed: 0	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won	avg_training_score	is_promoted
57564	52690	1	Analytics	region_7	Bachelor's	m	sourcing	2	29	3.0	5	0	85.0	0
11191	10257	2	Finance	region_2	Master's & above	f	sourcing	1	35	1.0	2	0	63.0	0
35941	32895	4	Sales & Marketing	region_2	Bachelor's	m	other	1	25	3.0	2	0	53.0	0
4809	4424	5	Analytics	region_7	Master's & above	m	other	2	46	3.0	7	0	86.0	0
45094	41261	7	Operations	region_32	Bachelor's	m	other	1	31	3.0	7	0	59.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
33930	31061	78292	Operations	region_2	Master's & above	m	sourcing	1	59	2.0	16	0	57.0	1
20809	19088	78294	Sales & Marketing	region_22	Bachelor's	m	sourcing	3	35	3.0	3	0	49.0	0
57590	52714	78296	Procurement	region_2	Bachelor's	f	sourcing	1	28	5.0	5	0	70.0	0
9855	9030	78297	Operations	region_13	Bachelor's	f	sourcing	1	34	5.0	7	0	56.0	0
57519	52646	78298	Procurement	region_2	Master's & above	f	referred	1	39	5.0	10	0	67.0	0

59858 rows × 14 columns

LINK GOOGLE COLAB SCRIPT

https://colab.research.google.com/drive/1L2mmKbIEgjifwAuVSrE1fPe8E1i2HsD3?usp=sharing

DigitalSkola

THANK YOU!

DigitalSkola