05_데이터 정리하기

데이터 정리하기

- 데이터 재구조화

> 데이터 정리하기

- melt() 메서드를 사용하여 열 데이터를 행으로 정리

```
billboard = pd.read_csv('data/billboard.csv')
billboard.head()
```

Out :

	year	artist	track	time	date.entered	wk1	wk2	wk3	wk4	wk5	...	wk67	wk68	wk69	wk70	wk71	wk72	wk73	wk74	wk75	wk76
0	2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26	87	82.0	72.0	77.0	87.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87.0	92.0	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70.0	68.0	67.0	66.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2000	3 Doors Down	Loser	4:24	2000-10-21	76	76.0	72.0	69.0	67.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	2000	504 Boyz	Wobble Wobble	3:35	2000-04-15	57	34.0	25.0	17.0	17.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

• 데이터 재구조화

> 데이터 정리하기

- melt() 메서드를 사용하여 열 데이터를 행으로 정리

```
billboard.columns
```

Out :

```
Index(['year', 'artist', 'track', 'time', 'date.entered', 'wk1', 'wk2', 'wk3',
       'wk4', 'wk5', 'wk6', 'wk7', 'wk8', 'wk9', 'wk10', 'wk11', 'wk12',
       'wk13', 'wk14', 'wk15', 'wk16', 'wk17', 'wk18', 'wk19', 'wk20', 'wk21',
       'wk22', 'wk23', 'wk24', 'wk25', 'wk26', 'wk27', 'wk28', 'wk29', 'wk30',
       'wk31', 'wk32', 'wk33', 'wk34', 'wk35', 'wk36', 'wk37', 'wk38', 'wk39',
       'wk40', 'wk41', 'wk42', 'wk43', 'wk44', 'wk45', 'wk46', 'wk47', 'wk48',
       'wk49', 'wk50', 'wk51', 'wk52', 'wk53', 'wk54', 'wk55', 'wk56', 'wk57',
       'wk58', 'wk59', 'wk60', 'wk61', 'wk62', 'wk63', 'wk64', 'wk65', 'wk66',
       'wk67', 'wk68', 'wk69', 'wk70', 'wk71', 'wk72', 'wk73', 'wk74', 'wk75',
       'wk76'],
      dtype='object')
```

데이터 정리하기

- 데이터 재구조화

04

> 데이터 정리하기
  - melt() 메서드를 사용하여 열 데이터를 행으로 정리

```
billboard.isnull().sum()
```

Out :

```
year            0
artist          0
track           0
time            0
date.entered    0
               ...
wk72          317
wk73          317
wk74          317
wk75          317
wk76          317
Length: 81, dtype: int64
```

```
billboard.dropna()
```

Out :

year	artist	track	time	date.entered	wk1	wk2	wk3	wk4	wk5	...	wk67	wk68	wk69	wk70	wk71	wk72	wk73	wk74	wk75	wk76

0 rows × 81 columns

데이터 정리하기

- 데이터 재구조화

> 데이터 정리하기

- melt() 메서드를 사용하여 열 데이터를 행으로 정리

```
b_long = pd.melt(billboard, id_vars = ['year', 'artist', 'track', 'time', 'date.entered'],
                 var_name = 'week', value_name = 'rating')
b_long
```

Out :

	year	artist	track	time	date.entered	week	rating
0	2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26	wk1	87.0
1	2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	wk1	91.0
2	2000	3 Doors Down	Kryptonite	3:53	2000-04-08	wk1	81.0
3	2000	3 Doors Down	Loser	4:24	2000-10-21	wk1	76.0
4	2000	504 Boyz	Wobble Wobble	3:35	2000-04-15	wk1	57.0
...	...	...	...	...	...	...	...
24087	2000	Yankee Grey	Another Nine Minutes	3:10	2000-04-29	wk76	NaN
24088	2000	Yearwood, Trisha	Real Live Woman	3:55	2000-04-01	wk76	NaN
24089	2000	Ying Yang Twins	Whistle While You Tw...	4:19	2000-03-18	wk76	NaN
24090	2000	Zombie Nation	Kernkraft 400	3:30	2000-09-02	wk76	NaN
24091	2000	matchbox twenty	Bent	4:12	2000-04-29	wk76	NaN

• 데이터 재구조화

> 데이터 정리하기

- melt() 메서드를 사용하여 열 데이터를 행으로 정리

```
b_long.dropna()
```

Out :

	year	artist	track	time	date.entered	week	rating
0	2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26	wk1	87.0
1	2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	wk1	91.0
2	2000	3 Doors Down	Kryptonite	3:53	2000-04-08	wk1	81.0
3	2000	3 Doors Down	Loser	4:24	2000-10-21	wk1	76.0
4	2000	504 Boyz	Wobble Wobble	3:35	2000-04-15	wk1	57.0
...	...	...	...	...	...	...	...
19716	2000	Creed	Higher	5:16	1999-09-11	wk63	50.0
19833	2000	Lonestar	Amazed	4:25	1999-06-05	wk63	45.0
20033	2000	Creed	Higher	5:16	1999-09-11	wk64	50.0
20150	2000	Lonestar	Amazed	4:25	1999-06-05	wk64	50.0
20350	2000	Creed	Higher	5:16	1999-09-11	wk65	49.0

5307 rows × 7 columns

• 데이터 재구조화

> 데이터 정리하기

- ebola 데이터 처리

```
ebola = pd.read_csv('data/country_timeseries.csv')
ebola.columns
```

Out :

```
Index(['Date', 'Day', 'Cases_Guinea', 'Cases_Liberia', 'Cases_SierraLeone',
       'Cases_Nigeria', 'Cases_Senegal', 'Cases_UnitedStates', 'Cases_Spain',
       'Cases_Mali', 'Deaths_Guinea', 'Deaths_Liberia', 'Deaths_SierraLeone',
       'Deaths_Nigeria', 'Deaths_Senegal', 'Deaths_UnitedStates',
       'Deaths_Spain', 'Deaths_Mali'],
      dtype='object')
```

• 데이터 재구조화

> 데이터 정리하기

- ebola 데이터 처리

```
e_long = pd.melt(ebola, id_vars = ['Date', 'Day'], value_name = 'count')
e_long.head()
```

Out :

	Date	Day	variable	count
0	1/5/2015	289	Cases_Guinea	2776.0
1	1/4/2015	288	Cases_Guinea	2775.0
2	1/3/2015	287	Cases_Guinea	2769.0
3	1/2/2015	286	Cases_Guinea	NaN
4	12/31/2014	284	Cases_Guinea	2730.0

• 데이터 재구조화

> 데이터 정리하기

- ebola 데이터 처리

```
e_long['status'] = e_long['variable'].apply(lambda x : x.split('_')[0])
e_long['country'] = e_long['variable'].apply(lambda x : x.split('_')[1])
e_long
```

Out :

	Date	Day	variable	count	status	country
0	1/5/2015	289	Cases_Guinea	2776.0	Cases	Guinea
1	1/4/2015	288	Cases_Guinea	2775.0	Cases	Guinea
2	1/3/2015	287	Cases_Guinea	2769.0	Cases	Guinea
3	1/2/2015	286	Cases_Guinea	NaN	Cases	Guinea
4	12/31/2014	284	Cases_Guinea	2730.0	Cases	Guinea
...	...	...	...	...	...	...
1947	3/27/2014	5	Deaths_Mali	NaN	Deaths	Mali
1948	3/26/2014	4	Deaths_Mali	NaN	Deaths	Mali
1949	3/25/2014	3	Deaths_Mali	NaN	Deaths	Mali
1950	3/24/2014	2	Deaths_Mali	NaN	Deaths	Mali
1951	3/22/2014	0	Deaths_Mali	NaN	Deaths	Mali

데이터 정리하기

- 데이터 재구조화

> 데이터 정리하기

- pivot_table() 메서드를 사용하여 행 데이터를 열로 정리

```
ebola_pivot = pd.pivot_table(e_long, index = ['Date', 'Day', 'country'],
                             columns = 'status', values = 'value')
ebola_pivot.head()
```

Out :

		status	Cases	Deaths
Date	Day	country		
1/2/2015	286	Liberia	8157.0	3496.0
1/3/2015	287	Guinea	2769.0	1767.0
		Liberia	8166.0	3496.0
		SierraLeone	9722.0	2915.0
1/4/2015	288	Guinea	2775.0	1781.0

• 데이터 재구조화

> 데이터 정리하기

- pivot_table() 메서드를 사용하여 행 데이터를 열로 정리

```
ebola_pivot.reset_index()
```

Out :

	status	Date	Day	country	Cases	Deaths
0		1/2/2015	286	Liberia	8157.0	3496.0
1		1/3/2015	287	Guinea	2769.0	1767.0
2		1/3/2015	287	Liberia	8166.0	3496.0
3		1/3/2015	287	SierraLeone	9722.0	2915.0
4		1/4/2015	288	Guinea	2775.0	1781.0
...		...	...	...	...	...
370		9/7/2014	169	Liberia	2081.0	1137.0
371		9/7/2014	169	Nigeria	21.0	8.0
372		9/7/2014	169	Senegal	3.0	0.0
373		9/7/2014	169	SierraLeone	1424.0	524.0
374		9/9/2014	171	Liberia	2407.0	NaN

데이터 정리하기

- 데이터 재구조화

> **데이터 정리하기**
  - pivot() 메서드를 사용하여 행 데이터를 열로 정리

```
ebola_pivot = pd.pivot(e_long, index = ['Date', 'Day', 'country'],
                                columns = 'status', values = 'value')
ebola_pivot.head()
```

Out :

		status	Cases	Deaths
Date	Day	country		
1/2/2015	286	Guinea	NaN	NaN
		Liberia	8157.0	3496.0
		Mali	NaN	NaN
		Nigeria	NaN	NaN
		Senegal	NaN	NaN

• 데이터 재구조화

> 데이터 정리하기

- melt, pivot_table를 사용하여 정리

```
weather = pd.read_csv('data/weather.csv')
print(weather.columns)
weather
```

Out :
```
Index(['id', 'year', 'month', 'element', 'd1', 'd2', 'd3', 'd4', 'd5', 'd6',
       'd7', 'd8', 'd9', 'd10', 'd11', 'd12', 'd13', 'd14', 'd15', 'd16',
       'd17', 'd18', 'd19', 'd20', 'd21', 'd22', 'd23', 'd24', 'd25', 'd26',
       'd27', 'd28', 'd29', 'd30', 'd31'],
      dtype='object')
```

	id	year	month	element	d1	d2	d3	d4	d5	d6	...	d22	d23	d24	d25	d26	d27	d28	d29	d30	d31
0	MX17004	2010	1	tmax	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	27.8	NaN
1	MX17004	2010	1	tmin	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	14.5	NaN
2	MX17004	2010	2	tmax	NaN	27.3	24.1	NaN	NaN	NaN	...	NaN	29.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	MX17004	2010	2	tmin	NaN	14.4	14.4	NaN	NaN	NaN	...	NaN	10.7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	MX17004	2010	3	tmax	NaN	NaN	NaN	NaN	32.1	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

• 데이터 재구조화

> 데이터 정리하기

- melt, pivot_table를 사용하여 정리

```
weather = pd.melt(weather, id_vars = ['id', 'year', 'month', 'element'],
                  var_name = 'day', value_name = 'temp')
weather
```

Out :

	id	year	month	element	day	temp
0	MX17004	2010	1	tmax	d1	NaN
1	MX17004	2010	1	tmin	d1	NaN
2	MX17004	2010	2	tmax	d1	NaN
3	MX17004	2010	2	tmin	d1	NaN
4	MX17004	2010	3	tmax	d1	NaN
...	...	...	...	...	...	...
677	MX17004	2010	10	tmin	d31	NaN
678	MX17004	2010	11	tmax	d31	NaN
679	MX17004	2010	11	tmin	d31	NaN
680	MX17004	2010	12	tmax	d31	NaN
681	MX17004	2010	12	tmin	d31	NaN

• 데이터 재구조화

> 데이터 정리하기

- melt, pivot_table를 사용하여 정리

```
weather['day'] = weather['day'].str.replace('d', '').astype(int)
weather
```

Out :

	id	year	month	element	day	temp
0	MX17004	2010	1	tmax	1	NaN
1	MX17004	2010	1	tmin	1	NaN
2	MX17004	2010	2	tmax	1	NaN
3	MX17004	2010	2	tmin	1	NaN
4	MX17004	2010	3	tmax	1	NaN
...	...	...	...	...	...	...
677	MX17004	2010	10	tmin	31	NaN
678	MX17004	2010	11	tmax	31	NaN
679	MX17004	2010	11	tmin	31	NaN
680	MX17004	2010	12	tmax	31	NaN
681	MX17004	2010	12	tmin	31	NaN

• 데이터 재구조화

> 데이터 정리하기

- melt, pivot_table를 사용하여 정리

```
weather_pivot = pd.pivot_table(weather, index = ['id', 'year', 'month', 'day'],
                               columns = 'element', values = 'temp')
weather_pivot
```

Out :

				element	tmax	tmin
id	year	month	day			
MX17004	2010	1	30		27.8	14.5
		2	2		27.3	14.4
			3		24.1	14.4
			11		29.7	13.4
			23		29.9	10.7
		3	5		32.1	14.2
			10		34.5	16.8
			16		31.1	17.6

• 데이터 재구조화

> 데이터 정리하기

- melt, pivot_table를 사용하여 정리

```
weather_pivot.reset_index()
```

Out :

element	id	year	month	day	tmax	tmin
0	MX17004	2010	1	30	27.8	14.5
1	MX17004	2010	2	2	27.3	14.4
2	MX17004	2010	2	3	24.1	14.4
3	MX17004	2010	2	11	29.7	13.4
4	MX17004	2010	2	23	29.9	10.7
5	MX17004	2010	3	5	32.1	14.2
6	MX17004	2010	3	10	34.5	16.8
7	MX17004	2010	3	16	31.1	17.6
8	MX17004	2010	4	27	36.3	16.7
9	MX17004	2010	5	27	33.2	18.2
10	MX17004	2010	6	17	28.0	17.5
11	MX17004	2010	6	29	30.1	18.0

• 중복 데이터 처리하기

> 빌보드 차트 중복데이터 처리
  - drop_duplicates() 메서드

```
billboard = pd.read_csv('data/billboard.csv')
b_long = pd.melt(billboard, id_vars = ['year', 'artist', 'track', 'time', 'date.entered'],
                 var_name = 'week', value_name = 'rating')
b_long.head()
```

Out :

	year	artist	track	time	date.entered	week	rating
0	2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26	wk1	87.0
1	2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	wk1	91.0
2	2000	3 Doors Down	Kryptonite	3:53	2000-04-08	wk1	81.0
3	2000	3 Doors Down	Loser	4:24	2000-10-21	wk1	76.0
4	2000	504 Boyz	Wobble Wobble	3:35	2000-04-15	wk1	57.0

• 중복 데이터 처리하기

> 빌보드 차트 중복데이터 처리
  - drop_duplicates() 메서드

```
b_long[b_long['track'] == 'Loser']
```

Out :

	year	artist	track	time	date.entered	week	rating
3	2000	3 Doors Down	Loser	4:24	2000-10-21	wk1	76.0
320	2000	3 Doors Down	Loser	4:24	2000-10-21	wk2	76.0
637	2000	3 Doors Down	Loser	4:24	2000-10-21	wk3	72.0
954	2000	3 Doors Down	Loser	4:24	2000-10-21	wk4	69.0
1271	2000	3 Doors Down	Loser	4:24	2000-10-21	wk5	67.0
...	...	...	...	...	...	...	...
22510	2000	3 Doors Down	Loser	4:24	2000-10-21	wk72	NaN
22827	2000	3 Doors Down	Loser	4:24	2000-10-21	wk73	NaN
23144	2000	3 Doors Down	Loser	4:24	2000-10-21	wk74	NaN
23461	2000	3 Doors Down	Loser	4:24	2000-10-21	wk75	NaN
23778	2000	3 Doors Down	Loser	4:24	2000-10-21	wk76	NaN

• 중복 데이터 처리하기

> 빌보드 차트 중복데이터 처리
  - drop_duplicates() 메서드

```
songs = b_long[['year', 'artist', 'track', 'time', 'date.entered']]
songs
```

Out :

	year	artist	track	time	date.entered
0	2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26
1	2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02
2	2000	3 Doors Down	Kryptonite	3:53	2000-04-08
3	2000	3 Doors Down	Loser	4:24	2000-10-21
4	2000	504 Boyz	Wobble Wobble	3:35	2000-04-15
...	...	...	...	...	...
24087	2000	Yankee Grey	Another Nine Minutes	3:10	2000-04-29
24088	2000	Yearwood, Trisha	Real Live Woman	3:55	2000-04-01
24089	2000	Ying Yang Twins	Whistle While You Tw...	4:19	2000-03-18
24090	2000	Zombie Nation	Kernkraft 400	3:30	2000-09-02
24091	2000	matchbox twenty	Bent	4:12	2000-04-29

• 중복 데이터 처리하기

> 빌보드 차트 중복데이터 처리
  - drop_duplicates() 메서드

```
songs.drop_duplicates()
```

Out :

	year	artist	track	time	date.entered
0	2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26
1	2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02
2	2000	3 Doors Down	Kryptonite	3:53	2000-04-08
3	2000	3 Doors Down	Loser	4:24	2000-10-21
4	2000	504 Boyz	Wobble Wobble	3:35	2000-04-15
...	...	...	...	...	...
312	2000	Yankee Grey	Another Nine Minutes	3:10	2000-04-29
313	2000	Yearwood, Trisha	Real Live Woman	3:55	2000-04-01
314	2000	Ying Yang Twins	Whistle While You Tw...	4:19	2000-03-18
315	2000	Zombie Nation	Kernkraft 400	3:30	2000-09-02
316	2000	matchbox twenty	Bent	4:12	2000-04-29

• 중복 데이터 처리하기

> 빌보드 차트 중복데이터 처리
  - drop_duplicates() 메서드

```
b_long.drop_duplicates(['year', 'artist', 'track', 'time', 'date.entered'])
```

Out :

	year	artist	track	time	date.entered	week	rating
0	2000	2 Pac	Baby Don't Cry (Keep...	4:22	2000-02-26	wk1	87.0
1	2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	wk1	91.0
2	2000	3 Doors Down	Kryptonite	3:53	2000-04-08	wk1	81.0
3	2000	3 Doors Down	Loser	4:24	2000-10-21	wk1	76.0
4	2000	504 Boyz	Wobble Wobble	3:35	2000-04-15	wk1	57.0
...	...	...	...	...	...	...	...
312	2000	Yankee Grey	Another Nine Minutes	3:10	2000-04-29	wk1	86.0
313	2000	Yearwood, Trisha	Real Live Woman	3:55	2000-04-01	wk1	85.0
314	2000	Ying Yang Twins	Whistle While You Tw...	4:19	2000-03-18	wk1	95.0
315	2000	Zombie Nation	Kernkraft 400	3:30	2000-09-02	wk1	99.0
316	2000	matchbox twenty	Bent	4:12	2000-04-29	wk1	60.0