Библиотека Pandas

Когда и зачем использовать

Режим презентации jupyter nbconvert my_script.ipynb --to slides --post serve

https://medium.com/@mjspeck/presenting-code-using-jupyter-notebook-slides-a8a3c3b59d67 (https://medium.com/@mjspeck/presenting-code-using-jupyter-notebook-slides-a8a3c3b59d67)

In [5]:

```
# запуск презентации
!/Users/kbashevoy/anaconda3/bin/jupyter nbconvert Python_10_pandas_presentation_mode.ipynb
```

```
[NbConvertApp] Converting notebook Python_10_pandas_presentation_mode.ipynb
to slides
[NbConvertApp] Writing 298005 bytes to Python_10_pandas_presentation_mode.sl
ides.html
[NbConvertApp] Redirecting reveal.js requests to https://cdnjs.cloudflare.co
m/ajax/libs/reveal.js/3.5.0 (https://cdnjs.cloudflare.com/ajax/libs/reveal.j
s/3.5.0)
Serving your slides at http://127.0.0.1:8000/Python_10_pandas_presentation_m
ode.slides.html (http://127.0.0.1:8000/Python_10_pandas_presentation_mode.sl
ides.html)
Use Control-C to stop this server
WARNING:tornado.access:404 GET /custom.css (127.0.0.1) 0.88ms
WARNING:tornado.access:404 GET /custom.css (127.0.0.1) 1.28ms
^C

Interrupted
```

In [ ]:

```
import pandas as pd
```

In [2]:

```
# https://grouplens.org/datasets/movielens/

ratings = pd.read_csv('ml-latest/ratings.csv')
ratings.head()
```

Out[2]:

	userId	movieId	rating	timestamp
0	1	110	1.0	1425941529
1	1	147	4.5	1425942435
2	1	858	5.0	1425941523
3	1	1221	5.0	1425941546
4	1	1246	5.0	1425941556

In [3]:

```
ratings.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26024289 entries, 0 to 26024288
Data columns (total 4 columns):
userId       int64
movieId      int64
rating       float64
timestamp    int64
dtypes: float64(1), int64(3)
memory usage: 794.2 MB
```

In [4]:

```
# сколько фильмов в коллекции
len(ratings.movieId.unique())
```

Out[4]:

```
45115
```

In [5]:

```
%%time

# топ пользователей по количеству оценок
ratings.groupby('userId').count().sort_values('movieId', ascending=False).head()
```

```
CPU times: user 1.05 s, sys: 693 ms, total: 1.74 s
Wall time: 1.83 s
```

Out[5]:

userId	movieId	rating	timestamp
45811	18276	18276	18276
8659	9279	9279	9279
270123	7638	7638	7638
179792	7515	7515	7515
228291	7410	7410	7410

Lifetime

In [6]:

```
ratings_grouped = ratings.groupby('userId').agg([min, max])
ratings_grouped.head()
```

Out[6]:

	movieId		rating		timestamp	
	min	max	min	max	min	max
userId						
1	110	112552	0.5	5.0	1425941300	1425942699
2	5	1552	1.0	5.0	867039165	867041296
3	480	4474	2.0	4.0	1048076830	1048077048
4	223	5679	1.0	5.0	1042667845	1042674886
5	7	3255	1.0	5.0	949423787	949424522

In [7]:

```
ratings_grouped['diff'] = ratings_grouped['timestamp']['max'] - ratings_grouped['timestamp'
ratings_grouped.head()
```

Out[7]:

	movieId		rating		timestamp		diff
	min	max	min	max	min	max	
userId							
1	110	112552	0.5	5.0	1425941300	1425942699	1399
2	5	1552	1.0	5.0	867039165	867041296	2131
3	480	4474	2.0	4.0	1048076830	1048077048	218
4	223	5679	1.0	5.0	1042667845	1042674886	7041
5	7	3255	1.0	5.0	949423787	949424522	735

In [8]:

```
ratings_grouped['diff'].mean() / 24 / 3600
```

Out[8]:

```
126.22761148027357
```

In [9]:

```
%%time

ratings = pd.read_csv('ml-latest/ratings.csv')
ratings_grouped = ratings.groupby('userId').agg([min, max])
ratings_grouped['diff'] = ratings_grouped['timestamp']['max'] - ratings_grouped['timestamp'

ratings_grouped['diff'].mean() / 24 / 3600
```

```
CPU times: user 14.3 s, sys: 2.41 s, total: 16.7 s
Wall time: 18 s
```

JOIN как в SQL

In [10]:

```
movies = pd.read_csv('ml-latest/movies.csv')
movies.head()
```

Out[10]:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

In [11]:

```
joined = ratings.merge(movies, on='movieId', how='left')
joined.head()
```

Out[11]:

	userId	movieId	rating	timestamp	title	genres
0	1	110	1.0	1425941529	Braveheart (1995)	Action|Drama|War
1	1	147	4.5	1425942435	Basketball Diaries, The (1995)	Drama
2	1	858	5.0	1425941523	Godfather, The (1972)	Crime|Drama
3	1	1221	5.0	1425941546	Godfather: Part II, The (1974)	Crime|Drama
4	1	1246	5.0	1425941556	Dead Poets Society (1989)	Drama