

Student ID: DH20042 Name : 弗兰克

1. The difference in movies rating by males and females

```
In [1]: import pandas as pd
import numpy as np
```

```
In [2]: # Load dataset of movielens
unames = ['user id', 'age', 'gender', 'occupation', 'zip code']
users = pd.read_csv('u.user', sep = '|', names = unames)
rnames = ['user id', 'item id', 'rating', 'timestamp']
ratings = pd.read_csv('u.data', sep = '\t', names = rnames)
```

```
In [4]: # Merge data
users_df = users.loc[:, ['user id', 'gender']]
ratings_df = ratings.loc[:, ['user id', 'rating']]
rating_df = pd.merge(users_df, ratings_df)
rating_df
```

Out[4]:

	user id	gender	rating
0	1	M	4
1	1	M	3
2	1	M	4
3	1	M	4
4	1	M	4
...
99995	943	M	1
99996	943	M	4
99997	943	M	3
99998	943	M	4
99999	943	M	2

100000 rows × 3 columns

```
In [5]: rating_df.groupby('gender').rating.std()
```

```
Out[5]: gender
F      1.170951
M      1.109556
Name: rating, dtype: float64
```

```
In [6]: rating_df.groupby('gender').rating.apply(pd.Series.std)
```

```
Out[6]: gender
F      1.170951
M      1.109556
Name: rating, dtype: float64
```

```
In [7]: rating_df.groupby(['user id', 'gender']).apply(np.mean)
```

```
/opt/anaconda3/lib/python3.9/site-packages/numpy/core/fromnumeric.py:337
0: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
    return mean(axis=axis, dtype=dtype, out=out, **kwargs)
```

```
Out[7]:
```

		user id	rating
user id	gender		
1	M	1.0	3.610294
2	F	2.0	3.709677
3	M	3.0	2.796296
4	M	4.0	4.333333
5	F	5.0	2.874286
...
939	F	939.0	4.265306
940	M	940.0	3.457944
941	M	941.0	4.045455
942	F	942.0	4.265823
943	M	943.0	3.410714

943 rows × 2 columns

```
In [8]: #Save the grouping
df1 = rating_df.groupby(['user id', 'gender']).apply(np.mean)
```

```
/opt/anaconda3/lib/python3.9/site-packages/numpy/core/fromnumeric.py:337
0: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise TypeError. Select only valid columns before calling the reduction.
    return mean(axis=axis, dtype=dtype, out=out, **kwargs)
```

```
In [9]: # Grouping by gender
df1.groupby('gender').rating.std()
```

```
Out[9]: gender
F      0.481241
M      0.430076
Name: rating, dtype: float64
```

```
In [10]: pd.pivot_table(df1, values = 'rating', index = 'gender', aggfunc = pd.Series
```

```
Out[10]:
```

	rating
gender	
F	0.481241
M	0.430076

```
In [12]: pd.pivot_table(rating_df, index = ['user id', 'gender'], values = 'rating')
```

```
Out[12]:
```

		rating
user id	gender	
1	M	3.610294
2	F	3.709677
3	M	2.796296
4	M	4.333333
5	F	2.874286
...
939	F	4.265306
940	M	3.457944
941	M	4.045455
942	F	4.265823
943	M	3.410714

943 rows x 1 columns

```
In [13]: t = pd.pivot_table(rating_df, index = ['user id', 'gender'], values = 'rating')
female = t.query("gender == ['F']")
pd.Series.std(female)
```

```
Out[13]: rating    0.481241
dtype: float64
```

```
In [15]: p = pd.pivot_table(rating_df, index = ['user id', 'gender'], values = 'rating')
male = t.query("gender == ['M']")
pd.Series.std(male)
```

```
Out[15]: rating    0.430076
dtype: float64
```

```
In [ ]:
```

