Exploring the MovieLens Dataset

April 15, 2021

[1]:
```
import pandas as pd
```

[2]:
```
# import each of the three tables and assign names to each of the columns
unames = ['user_id', 'gender', 'age', 'occupation', 'zip']
users = pd.read_table('ml-1m/users.dat', sep='::', header=None, names=unames,␣
 ↪engine='python')

rnames = ['user_id', 'movie_id', 'rating', 'timestamp']
ratings = pd.read_table('ml-1m/ratings.dat', sep='::', header=None,␣
 ↪names=rnames, engine='python')

mnames = ['movie_id', 'title', 'genres']
movies = pd.read_table('ml-1m/movies.dat', sep='::', header=None, names=mnames,␣
 ↪engine='python')
```

[3]:
```
# take a look at the first 5 rows of each table:
users[:5]
```

[3]:

	user_id	gender	age	occupation	zip
0	1	F	1	10	48067
1	2	M	56	16	70072
2	3	M	25	15	55117
3	4	M	45	7	02460
4	5	M	25	20	55455

[4]:
```
ratings[:5]
```

[4]:

	user_id	movie_id	rating	timestamp
0	1	1193	5	978300760
1	1	661	3	978302109
2	1	914	3	978301968
3	1	3408	4	978300275
4	1	2355	5	978824291

[5]:
```
movies[:5]
```

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

[6]:
```
# Merge tables
data = pd.merge(pd.merge(ratings, users), movies)
```

[7]:
```
# Show the first row of the database
data.head(1)
```

[7]:
	user_id	movie_id	rating	timestamp	gender	age	occupation	zip	title	genres
0	1	1193	5	978300760	F	1	10	48067	One Flew Over the Cuckoo's Nest (1975)	Drama

1 1. An aggregate on the number of rating done for each particular genre

[8]:
```
# get movies of each particular genre
part_ratings = data.set_index(data.columns.drop('genres',1).tolist()).genres.
 ↪str.split('|', expand=True).stack().reset_index().rename(columns={0:
 ↪'genres'}).loc[:, data.columns]
# get the aggregate on the number of rating for each particular genre
agg_num_genre_ratings = part_ratings.pivot_table('rating', index='genres',␣
 ↪aggfunc=len)
agg_num_genre_ratings
```

[8]:
genres	rating
Action	257457
Adventure	133953
Animation	43293
Children's	72186
Comedy	356580
Crime	79541
Documentary	7910
Drama	354529
Fantasy	36301
Film-Noir	18261
Horror	76386
Musical	41533
Mystery	40178

```
Romance      147523
Sci-Fi       157294
Thriller     189680
War           68527
Western       20683
```

2 2. The top 5 ranked genres by women on most number of rating.

[9]:
```
# aggregate number of rating of particular genres by gender
num_ratings_gender = part_ratings.pivot_table('rating', index='genres',␣
 ↪columns='gender', aggfunc=len)
# rank number of rating by women
top_female_ratings = num_ratings_gender.sort_values(by='F', ascending=False)
# display top 5 rankings
top_female_ratings[:5]
```

[9]:
```
gender        F       M
genres
Drama     98153  256376
Comedy    96271  260309
Romance   50297   97226
Action    45650  211807
Thriller  40308  149372
```

3 3. The top 5 ranked genres by men on most number of rating.

[10]:
```
# rank number of rating by men
top_male_ratings = num_ratings_gender.sort_values(by='M', ascending=False)
# display top 5 rankings
top_male_ratings[:5]
```

[10]:
```
gender        F       M
genres
Comedy    96271  260309
Drama     98153  256376
Action    45650  211807
Thriller  40308  149372
Sci-Fi    27400  129894
```

4 4. Provide average animation movie's ratings by the following four time intervals during which the movies were released (a) 1970 to 1979 (b) 1980 to1989 (c) 1990 to 1999 (d) 2000 to 2009.

[11]:
```
sp_data = part_ratings
# split the 'title' column of data to 'title' and 'released_year' columns
sp_data[['title', 'release_year']] = sp_data.title.str.rsplit("(",1,expand=True)
sp_data['release_year'] = sp_data.release_year.str.replace(")", "")
# get the animation genre of data
sp_movie_of_genre = sp_data[(sp_data['genres'] == 'Animation')]
# group the dataframe by 4 time intervals
labels = ['1970-1979', '1980-1989', '1990-1999', '2000-2009']
sp_movie_of_genre['release_year_range'] = pd.cut(sp_movie_of_genre.release_year.
 ↪astype(int), range(1969, 2010, 10), right=True, labels=labels)
# create a pivot table with genre, releas_year_range and average ratings of␣
 ↪movies in each time interval
avg_rating_of_genre_by_range = sp_movie_of_genre.pivot_table('rating',␣
 ↪index=['genres','release_year_range'], aggfunc='mean')
avg_rating_of_genre_by_range
```

```
<ipython-input-11-c9a5d7b1df38>:9: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  sp_movie_of_genre['release_year_range'] =
pd.cut(sp_movie_of_genre.release_year.astype(int), range(1969, 2010, 10),
right=True, labels=labels)
```

[11]:

		rating
genres	release_year_range	
Animation	1970-1979	3.540103
	1980-1989	3.591174
	1990-1999	3.743986
	2000-2009	3.388091

5 5. A function that given a genre and a rating_range(i.e. [3.5, 4]), returns all the movies of that genre and within that rating range sorted by average rating.

[12]:
```
def genre_of_movies_within_rating_range(genre, range):
    # get the movies of the given genre
    movie_of_genre = part_ratings[(part_ratings['genres'] == genre)]
    # create a pivot table with title, genres and average ratings of movies
```

```
    movie_avg_rating_of_genre = movie_of_genre.pivot_table('rating',␣
↪index=['title','genres'], aggfunc='mean')
    # convert pivot dable to normal dataframe
    df = movie_avg_rating_of_genre.reset_index()
    # filter the frame by range of rating
    df_within_rating_range = df.loc[(df['rating'].isin(range),['title',␣
↪'genres', 'rating'])]
    # sort the data frame by average rating
    sorted_movie_rating_within_range = df_within_rating_range.
↪sort_values(by='rating', ascending=False)
    return sorted_movie_rating_within_range

# print drama genre moviews within the rating range (sorted by average rating␣
↪descendingly)
print (genre_of_movies_within_rating_range('Drama', [3.5,4]))
```

```
                                                 title genres  rating
7                              24 7: Twenty Four Seven   Drama     4.0
858         Midaq Alley (Callej n de los milagros, El)   Drama     4.0
601                                  Heaven's Burning   Drama     4.0
638   I Don't Want to Talk About It (De eso no se ha…  Drama     4.0
42                                    Alley Cats, The   Drama     4.0
652                                           Illtown   Drama     4.0
686                                Jar, The (Khomreh)   Drama     4.0
755                           Leather Jacket Love Story   Drama     4.0
787                               Lonely Are the Brave   Drama     4.0
992                                     Outside Ozona   Drama     4.0
509                                             Fresh   Drama     4.0
1146                                  Running Free   Drama     4.0
1149            Sacco and Vanzetti (Sacco e Vanzetti)   Drama     4.0
1158                                          Savior   Drama     4.0
1294                                          Target   Drama     4.0
1299                                        Ten Benny   Drama     4.0
1404             Voyage to the Beginning of the World   Drama     4.0
1409                               Walk in the Sun, A   Drama     4.0
511                           Friend of the Deceased, A   Drama     4.0
1423                            Wend Kuuni (God's Gift)   Drama     4.0
224                                           Bye-Bye   Drama     4.0
340                                             Daens   Drama     4.0
264                                         Chushingura   Drama     4.0
62                                              Angela   Drama     4.0
454                                         Fall Time   Drama     4.0
65                                              Anna   Drama     4.0
440                                Everything Relative   Drama     4.0
401   Dreamlife of Angels, The (La Vie r v e des ang…  Drama     4.0
382                                             Dingo   Drama     4.0
69                                Another Man's Poison   Drama     4.0
```

```
128                        Beloved/Friend (Amigo/Amado)   Drama     4.0
298                                     Condition Red   Drama     4.0
160                                   Blood on the Sun   Drama     3.5
1168                                      Second Best   Drama     3.5
482                              First Love, Last Rites   Drama     3.5
1211                          Slaves to the Underground   Drama     3.5
231                                           Captives   Drama     3.5
1390                                          Urbania   Drama     3.5
1421                               Welcome To Sarajevo   Drama     3.5
1329               Tigrero: A Film That Was Never Made   Drama     3.5
931                                           Niagara   Drama     3.5
1144                           Run of the Country, The   Drama     3.5
1082                                    Raining Stones   Drama     3.5
463                              Farinelli: il castrato   Drama     3.5
273                                        City, The   Drama     3.5
722                      Killing of Sister George, The   Drama     3.5
364                              Death and the Maiden   Drama     3.5
678                  JLG/JLG - autoportrait de d cembre   Drama     3.5
654                                           Impact   Drama     3.5
648   Identification of a Woman (Identificazione di …  Drama     3.5
614                                        Hollow Reed   Drama     3.5
592                                    Hangmen Also Die   Drama     3.5
555                                    Golden Bowl, The   Drama     3.5
645                    I'll Never Forget What's 'is Name   Drama     3.5
```

6 6.Present the top 50 ranked movies by highest ratings to generate a watching list

[13]:
```
# The top 50 ranked movies by highest ratings to generate a watching list
# aggregate number of rating of particular genres
avg_ratings = data.pivot_table('rating', index=['title','genres'],␣
 ↪aggfunc="mean")
# rank rating from high to low
top_avg_ratings = avg_ratings.sort_values(by='rating', ascending=False)
top_avg_ratings.head(50)
```

[13]:
```
                                                    rating
title                                                      genres
Ulysses (Ulisse) (1954)                                    Adventure
5.000000
Lured (1947)                                               Crime
5.000000
Follow the Bitch (1998)                                    Comedy
5.000000
Bittersweet Motel (2000)                                   Documentary
5.000000
```

```
Song of Freedom (1936)                                  Drama
5.000000
One Little Indian (1973)                                Comedy|Drama|Western
5.000000
Smashing Time (1967)                                    Comedy
5.000000
Schlafes Bruder (Brother of Sleep) (1995)               Drama
5.000000
Gate of Heavenly Peace, The (1995)                      Documentary
5.000000
Baby, The (1973)                                        Horror
5.000000
I Am Cuba (Soy Cuba/Ya Kuba) (1964)                     Drama
4.800000
Lamerica (1994)                                         Drama
4.750000
Apple, The (Sib) (1998)                                 Drama
4.666667
Sanjuro (1962)                                          Action|Adventure
4.608696
Seven Samurai (The Magnificent Seven) (Shichini… Action|Drama
4.560510
Shawshank Redemption, The (1994)                        Drama
4.554558
Godfather, The (1972)                                   Action|Crime|Drama
4.524966
Close Shave, A (1995)                                   Animation|Comedy|Thriller
4.520548
Usual Suspects, The (1995)                              Crime|Thriller
4.517106
Schindler's List (1993)                                 Drama|War
4.510417
Wrong Trousers, The (1993)                              Animation|Comedy
4.507937
Dry Cleaning (Nettoyage  sec) (1997)                    Drama
4.500000
Inheritors, The (Die Siebtelbauern) (1998)              Drama
4.500000
Mamma Roma (1962)                                       Drama
4.500000
Bells, The (1926)                                       Crime|Drama
4.500000
Dangerous Game (1993)                                   Drama
4.500000
Hour of the Pig, The (1993)                             Drama|Mystery
4.500000
Callej n de los milagros, El (1995)                     Drama
```

```
4.500000
Skipped Parts (2000)                                    Drama|Romance
4.500000
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)           Film-Noir
4.491489
Raiders of the Lost Ark (1981)                          Action|Adventure
4.477725
Rear Window (1954)                                      Mystery|Thriller
4.476190
Paths of Glory (1957)                                   Drama|War
4.473913
Star Wars: Episode IV - A New Hope (1977)               Action|Adventure|Fantasy|Sci-
Fi  4.453694
Third Man, The (1949)                                   Mystery|Thriller
4.452083
Dr. Strangelove or: How I Learned to Stop Worry… Sci-Fi|War
4.449890
For All Mankind (1989)                                  Documentary
4.444444
Wallace & Gromit: The Best of Aardman Animation… Animation
4.426941
To Kill a Mockingbird (1962)                            Drama
4.425647
Double Indemnity (1944)                                 Crime|Film-Noir
4.415608
Casablanca (1942)                                       Drama|Romance|War
4.412822
World of Apu, The (Apur Sansar) (1959)                  Drama
4.410714
Sixth Sense, The (1999)                                 Thriller
4.406263
Yojimbo (1961)                                          Comedy|Drama|Western
4.404651
Pather Panchali (1955)                                  Drama
4.404255
Lawrence of Arabia (1962)                               Adventure|War
4.401925
Return with Honor (1998)                                Documentary
4.400000
Maltese Falcon, The (1941)                              Film-Noir|Mystery
4.395973
One Flew Over the Cuckoo's Nest (1975)                  Drama
4.390725
Citizen Kane (1941)                                     Drama
4.388889
```